AD-A163 920   INTRODUCTION TO KERNEL DENSITY ESTIMATION(U) ARMY   1/1
               BALLISTIC RESEARCH LAB ABERDEEN PROVING GROUND MD
   W A WINNER DEC 85 BRL-TR-2705
UNCLASSIFIED   F/G 12/1   NL

END
FILMED
LING
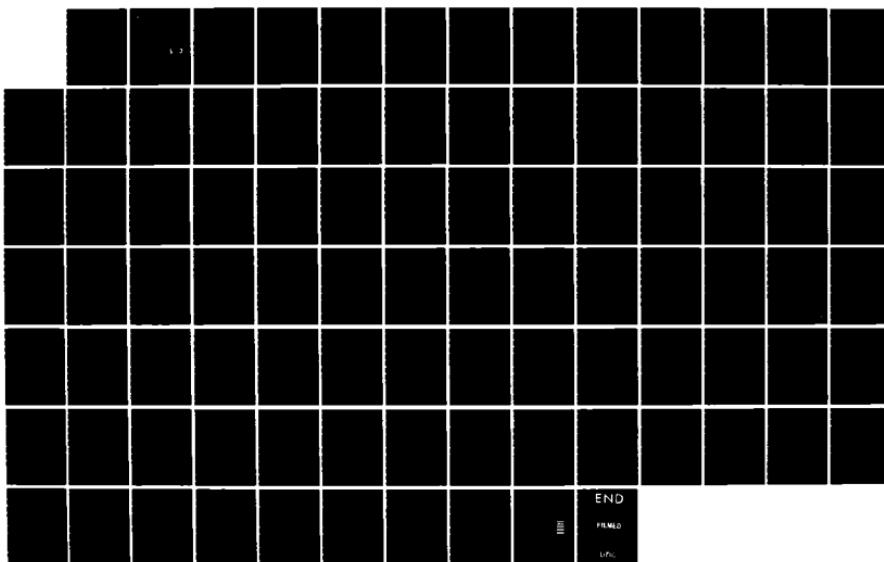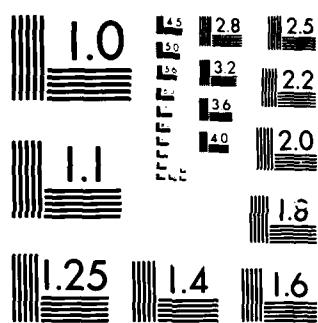
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

AD

US ARMY
MATERIEL
COMMAND

TECHNICAL REPORT BRL-TR-2705

AD-A163 920

INTRODUCTION TO KERNEL DENSITY ESTIMATION

Wendy A. Winner

December 1985

DTIC
ELECTE
FEB 05 1986
S
D
D

US ARMY BALLISTIC RESEARCH LABORATORY
ABERDEEN PROVING GROUND, MARYLAND

DTIC FILE COPY

86 2 5 060

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
Technical Report BRL-TR-2705	AD-A, 6392 o	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
Introduction to Kernel Density Estimation	TR FINAL
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Wendy A. Winner	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
US Army Ballistic Research Laboratory ATTN: SLCBR-SE Aberdeen Proving Ground, MD 21005-5066	

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
US Army Ballistic Research Laboratory ATTN: SLCBR-DD-T Aberdeen Proving Ground, MD 21005-5066	December 1985
	13. NUMBER OF PAGES
	100

14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)	15. SECURITY CLASS. (of this report)
	UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Probability Density Function
Kernel Density Estimation
Kernel Function
Smoothing Parameter
Fire Support Team (FIST)

Headquarters (HQ)
Force Development Testing and
   Experimentation II (FDT&E II)
Artillery Target Intelligence (ATI)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Kernel density estimators are one technique for producing nonparametric estimates of a sample's underlying probability density function. Although there are numerous kernel functions, the reader is only introduced to the Gaussian, truncated Gaussian, mode centering Lognormal and median centering Lognormal kernels. These kernels are applied to two samples from the Fire Support Team (FIST) Force Development Testing and Experimentation II (FDT&E II) conducted by the US Army Field Artillery Board (Fort Sill, OK), at Fort Riley, KS, during April and May 1984. Analysis of the performance of the Gaussian and

DD FORM 1473   EDITION OF 1 NOV 65 IS OBSOLETE

20. ABSTRACT (Cont.)

truncated Gaussion kernels is achieved by applying a recursive formula, developed by Richard A. Tapia and James R. Thompson, optimizing the kernel function's smoothing parameter for the FIST FDT&E II samples and for Monte Carlo simulations.

Keywords: Kernel functions

TABLE OF CONTENTS

3

TABLE OF CONTENTS (contd)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (contd)

LIST OF TABLES

I. INTRODUCTION

One technique for producing nonparametric estimates of a sample's underlying probability density function and its associated cumulative distribution function is kernel estimators. A kernel density estimate, $\hat{f}(z)$, is defined for a sample with values $x_1, x_2, \ldots, x_n$ at each fixed point z as follows:

$$\hat{f}(z) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{z - x_j}{h}\right)$$

where

$$\hat{f}(z) \geq 0 \text{ for all } z$$

$$\int_{-\infty}^{\infty} \hat{f}(z) dz = 1$$

where the kernel function K is a probability density function symmetric about zero and the data-based smoothing parameter h is strictly positive.[1] A kernel estimate of a sample's underlying cumulative distribution function, $\hat{F}(z)$, can be computed by integrating $\hat{f}(z)$ with the kernel function's proper limits of integration.

Kernel estimators, however, are plagued with problems ranging from choosing an appropriate kernel estimator, assessing the robustness of a kernel estimator and determining the optimum value for the kernel estimator's smoothing parameter given a particular sample. Despite these limitations kernel density estimators should not be overlooked among other nonparametric estimators such as histograms.

The purpose of this paper is to present an introduction to kernel estimators by examining the Gaussian and Lognormal kernels. These kernel estimators will be applied to two Artillery Target Intelligence message (ATIs) samples (**Tables 1** and **2**) from the Fire Support Team (FIST) Force Development Testing and Experimentation II (FDT&E II) conducted by the U.S. Army Field Artillery Board at Fort Riley, KS during April and May 1984.[2] Although a kernel density estimate is defined as a normalized density estimate, the figures in this report will depict unnormalized density estimates based on a size spacing, z, equal to one tenth of a second. Work with these samples suggested modifying both the Gaussian and Lognormal kernels. Subsequently, these modified kernels will also be presented. In addition, the previously mentioned problems associated with kernel estimation will be examined to clarify the magnitude of their influence and to show the usefulness of kernel density estimation.

---

[1]Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981, p 10

[2]Kaste, Virginia A., Brodeen, Douglas C. and Winner, Wendy A., "Description of the Digital Data Collected from the FIST FDT&E II," Ballistic Research Laboratory Technical Report No. 2676, 1985.

TABLE 1. FIST HQ Service Time
ATI Auto
FEX I and FEX II

Service Time (seconds)	Number of Occurrences
0	11
1	7
2	2
3	4
6	1
9	1
10	1
12	1
14	1
15	1
45	1

Total Number of Service Time Observations = 31

TABLE 2. FIST HQ Service Time
ATI Review
FEX I and FEX II

Service Time (seconds)	Number of Occurrences
0	1
1	2
2	3
6	1
7	2
16	1
18	1
27	1
33	1
35	1
69	1
112	1
280	1
444	1
831	1
850	1
867	1
2235	1

Total Number of Service Time Observations = 22

II. GAUSSIAN KERNEL

Definition

A Gaussian kernel estimate of a sample's underlying probability density function uses the normal probability density function for the kernel function and is defined as follows:[3]

For z: $z \in \mathbf{R}$, $-\infty < z < \infty$

$$\hat{f}(z) = \frac{1}{nh} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-1}{2} \left[\frac{z-x_j}{h} \right]^2 \right)}$$

Since the normal probability density function is a unimodal function symmetric about zero and defined for values ranging from $-\infty$ to ∞, this kernel density estimator can be applied to real-valued samples whose sample values range from $-\infty$ to ∞. In addition, this kernel will produce a density estimate that is symmetric about each sample value. Appendix A contains a user's guide for the FORTRAN programs provided in Appendices B through G. Appendix B contains the FORTRAN code programming the Gaussian kernel's estimate of a sample's underlying probability density function and cumulative distribution function.

Application of the Gaussian kernel to the FIST HQ samples

Figures 1 through 3 present Gaussian kernel density estimates for the ATI Auto sample superimposed on the sample's histogram showing the number of service time observations. Each of these estimates was produced by using different values for h, the data-based smoothing parameter. Figure 1 with h equal to 2.00 seems to oversmooth this FIST HQ service time sample. In this report, mode refers to the number of "local maxima" in a density estimate and differs from the statistical definition of mode(e.g., the observation(s) which occur(s) with the most frequency). Figure 2 with h equal to 1.00 produces a density estimate with five significant modes and seems to give an acceptable density estimate for this data sample. In Figure 3 with h equal to 0.50, the density estimate has seven modes and seems overly sensitive.

Figures 4 through 6 present Gaussian kernel density estimates for the ATI Review sample in the interval [0,50] for h equal to 5.00, 3.00 and 1.00, respectively. Both density estimates in Figures 4 and 5 seem to oversmooth this sample. However, the kernel density estimate in Figure 6 suggests five significant modes for the ATI Review sample in the interval [0.50] and seems to supply an acceptable density estimate for the ATI

---

[3]Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, p 61.

11
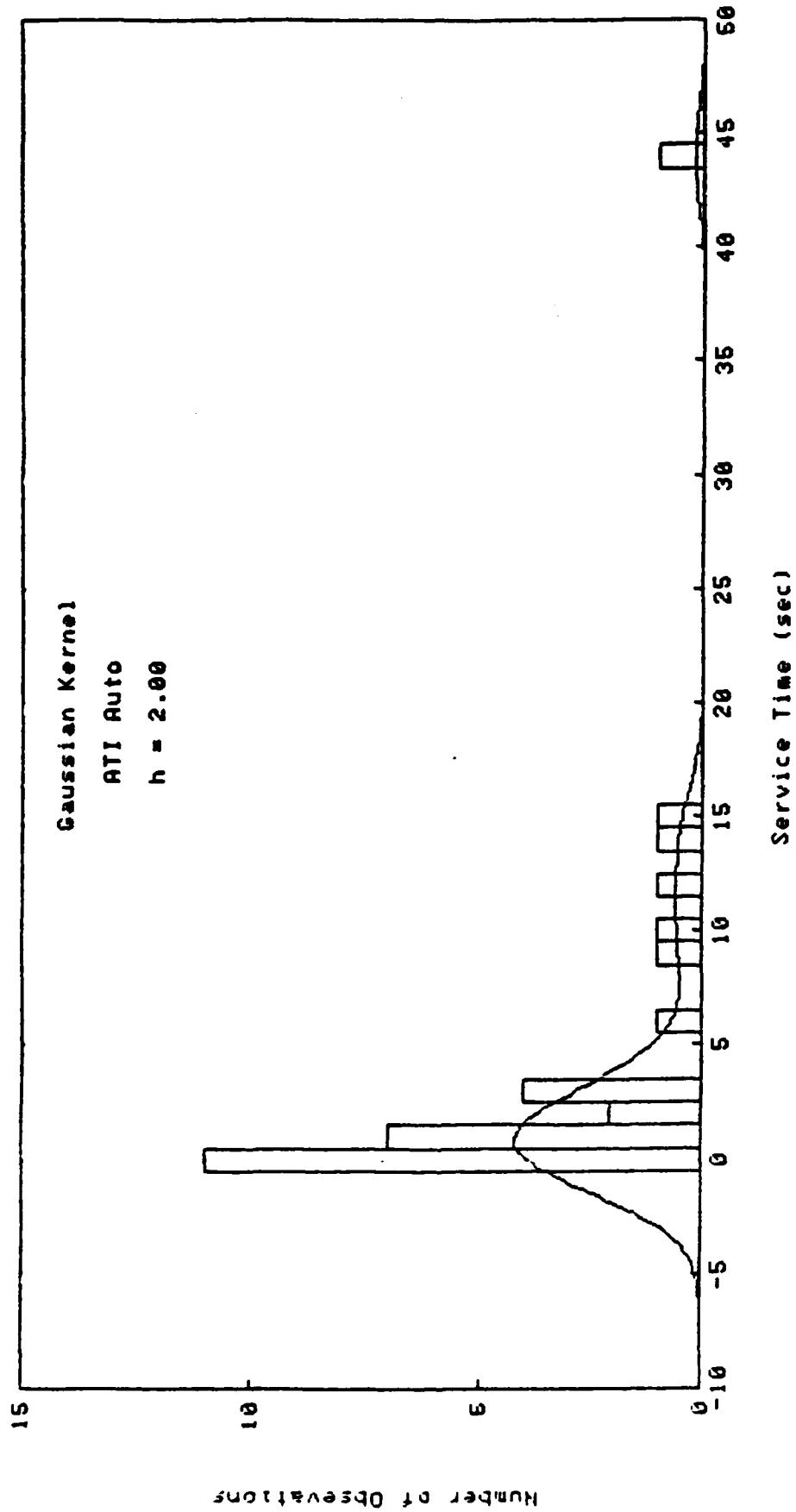
Figure 1. Application of the Gaussian Kernel with h=2.00 to ATI Auto Sample.
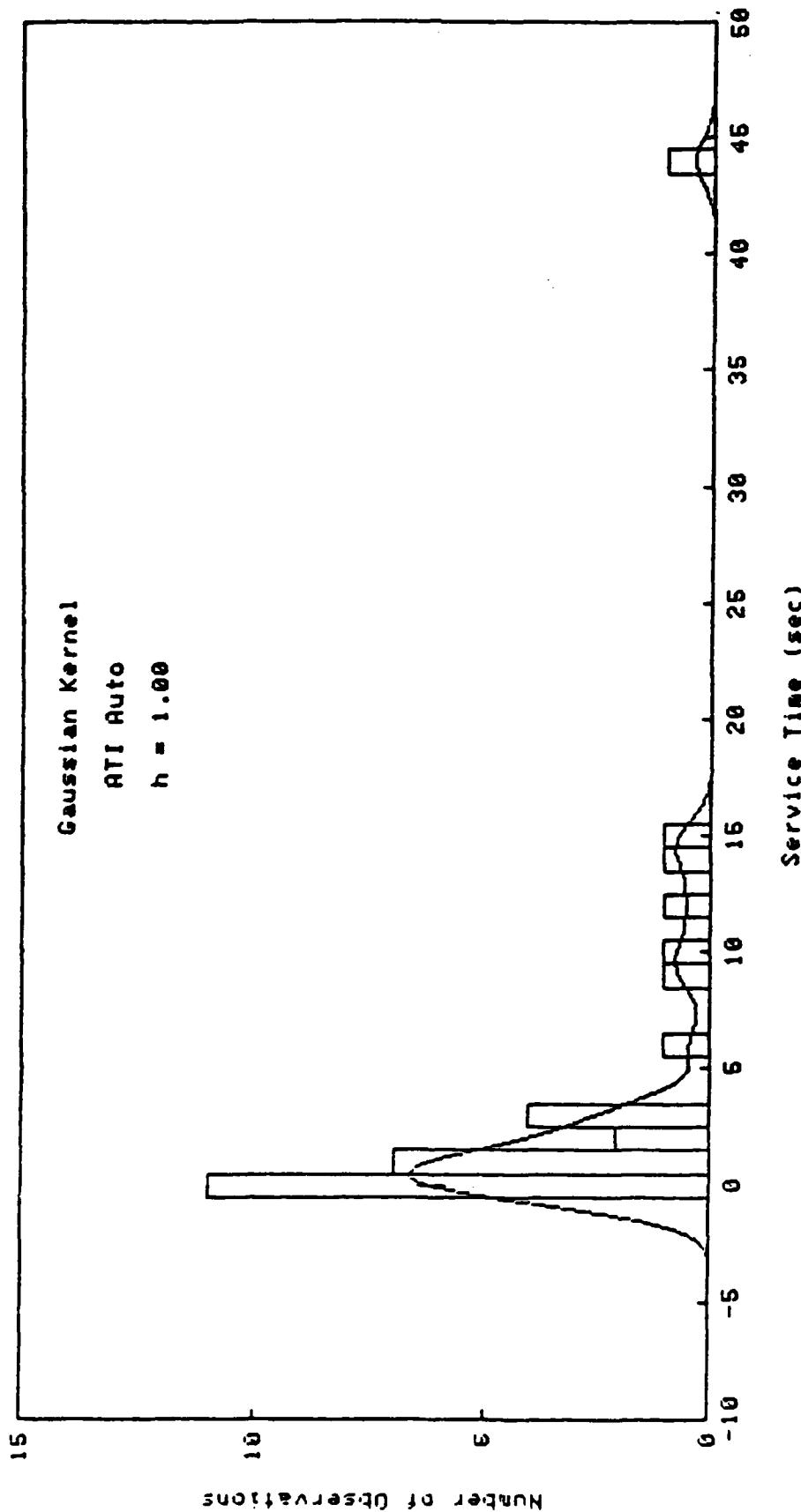
Figure 2. Application of the Gaussian Kernel with h=1.00 to ATI Auto Sample.

13
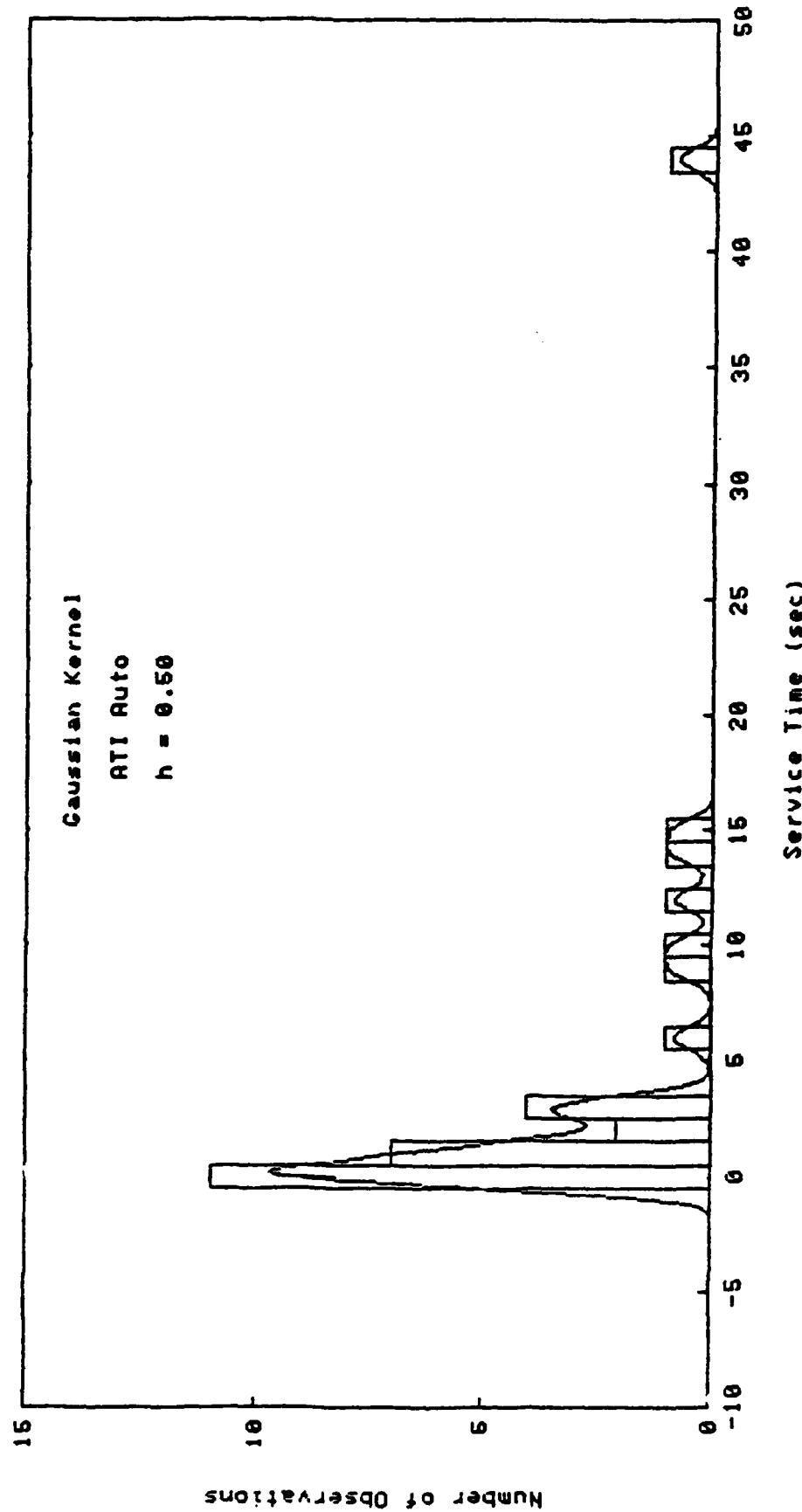
Figure 3. Application of the Gaussian Kernel with h=0.50 to ATl Auto Sample.

14
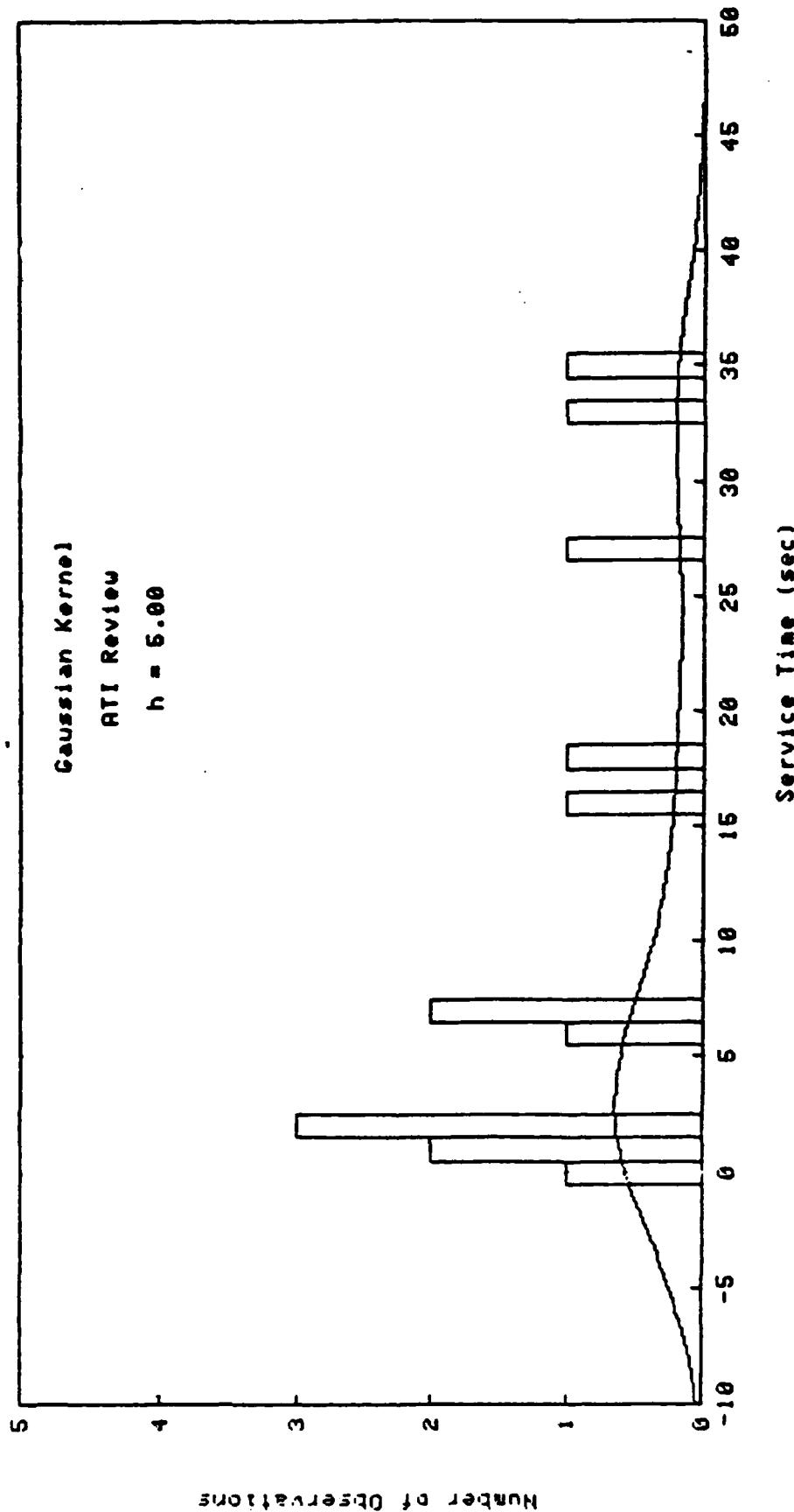
Figure 4. Application of the Gaussian Kernel with h=5.00 to ATI Review Sample.

15
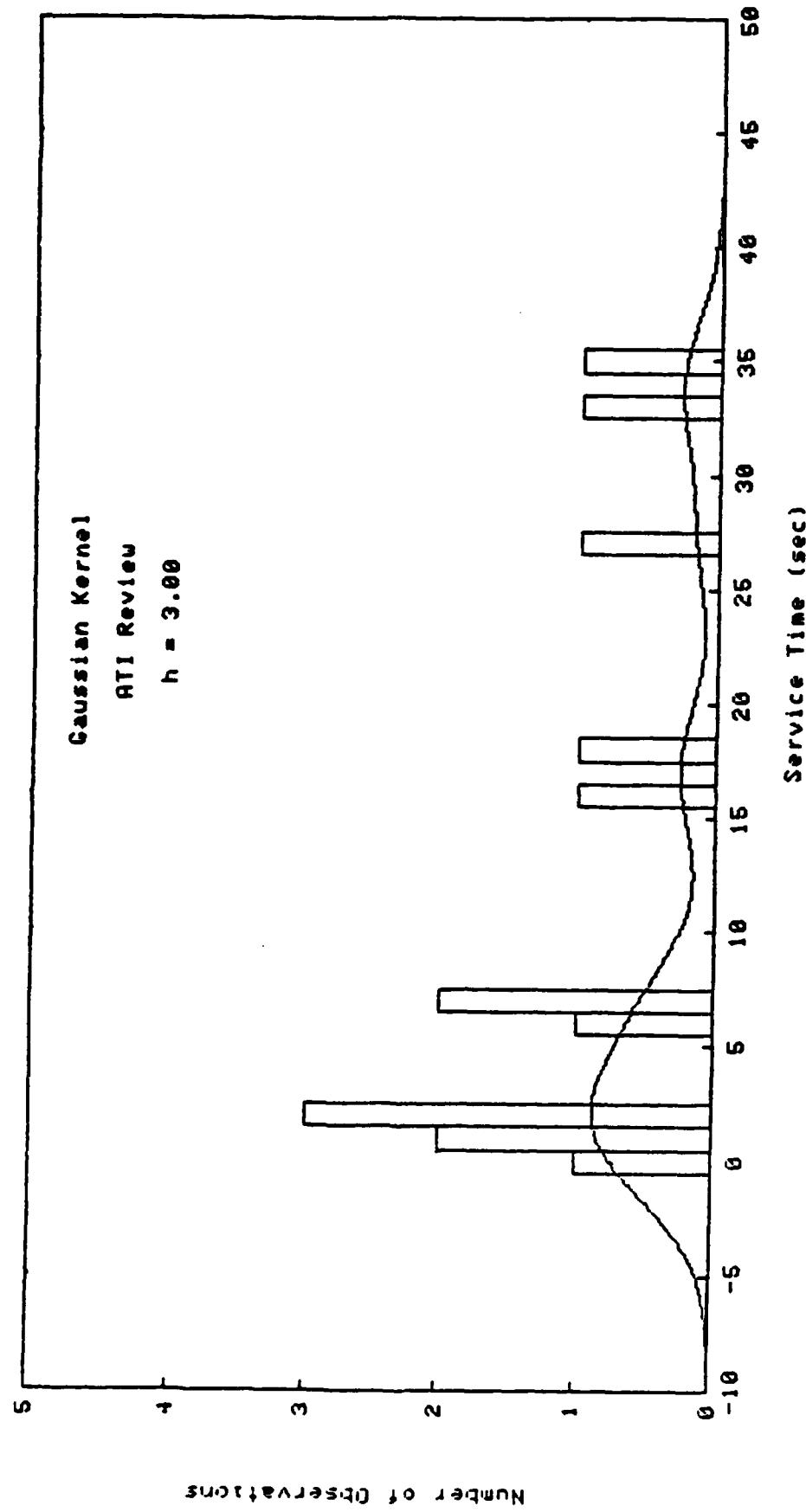
Figure 5. Application of the Gaussian Kernel with h=3.0u to ATI Review Sample.

16
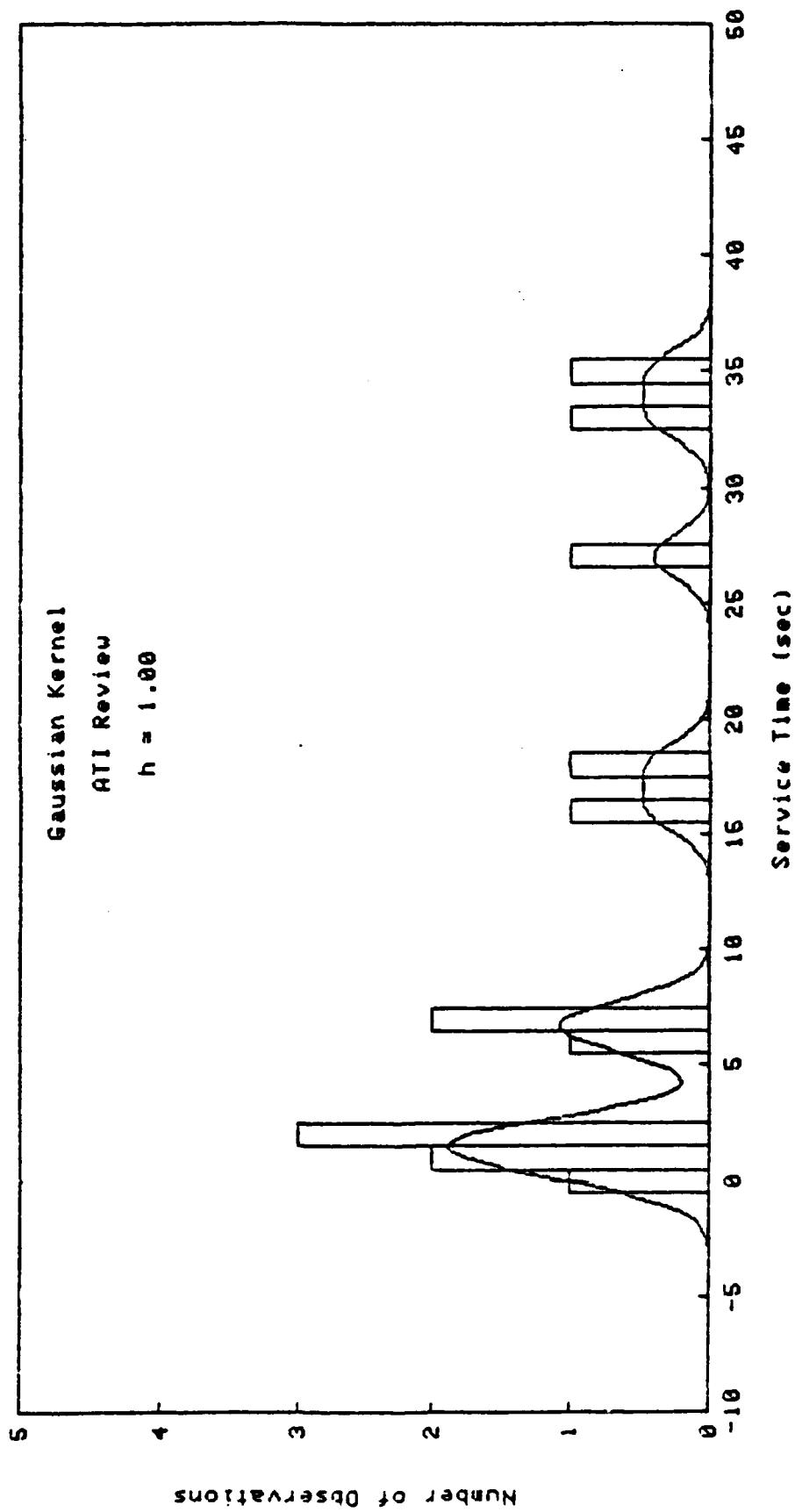
Figure 6. Application of the Gaussian Kernel with h=1.00 to ATI Review Sample.

17

Review sample.

From **Figures 1** through **6**, one can see that the value chosen for h varies the kernel density estimate. To understand the relationship between the h value and the number of modes in the estimate, **Table 3** presents the critical h values for the FIST HQ samples. A critical h value is the smallest h value producing a kernel density estimate with m modes, and therefore, h values smaller than the critical h value produce kernel density estimates with more than m modes. The critical h values in **Table 3** were obtained by interactively selecting h values that, in turn, were used to produce kernel density estimates.

TABLE 3. Gaussian Kernel's Critical h Values
for the FIST HQ Samples

Sample	Number of Modes in interval [0.50]	Critical h Value to two decimal places
ATI Auto	1	10.01
	2	2.73
	3	1.85
	4	1.01
	5	$1.00 < h < 1.01$
	6	0.65
	7	0.50
ATI Review	1	7.05
	2	4.08
	3	2.60
	4	1.94
	5	1.00
	6	$0.99 < h < 1.00$
	7	0.42

From **Table 3** one can see that as the h value increases the smoothness of the kernel density estimate increases. The role that h performs for kernel estimation is analogous to the role that the (data-based) interval size performs for a histogram. For a histogram the general rule of thumb for an appropriate interval size is 10 to 25 intervals covering the range of the sample.[4]

Tapia and Thompson have developed a recursive formula for determining the global optimal h for a kernel estimator by minimizing the integrated mean square error(IMSE). The IMSE and the recursive formula is defined as follows:[5]

[4]Guttman, I. and Wilks, S. S., *Introductory Engineering Statistics*, New York: John Wiley & Sons, 1965 cited by Scott, Da i d W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981, p. 9.

[5] Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, pp. 48,50.

$$\text{IMSE}[\hat{f}] = \int E\left[(\hat{f}(z) - f(z))^2\right]dz$$

$$\text{IMSE}[\hat{f}] \approx \frac{1}{nh}\left[\int [K(y)]^2 dy\right] + h^{(2r)}\left[\int \frac{y^r K(y)}{r!}dy\right]^2\left[\int \left(f^{(r)}(z)\right)^2 dz\right]$$

and

$$h^{i+1} = n^{\left(\frac{-1}{2r+1}\right)}\alpha(K)\beta(f)$$

where

$$\alpha(K) = \left\{\frac{\int K^2(y)dy}{2r\left(\int \frac{y^r K(y)}{r!}dy\right)^2}\right\}^{\frac{1}{2r+1}}$$

$$\beta(f) = \left(\int (f^{(r)}(z))^r dz\right)^{\frac{-1}{2r+1}}.$$

Thus, each successive value for h, h^{i+1}, is computed by using the previous h value, h^i, contained in $f^{(r)}(z)$. In application, $\hat{f}(z)$ estimates $f(z)$. The characteristic exponent, r, must be determined for each kernel function K, and Tapia and Thompson report that the Gaussian kernel's characteristic exponent is 2.[6]

The attractiveness of Tapia and Thompson's formula is that it is tolerant of extreme guesses for the initial h value. More importantly, however, Scott and Factor have noted that choosing the initial h value to be the sample's range guarantees the convergence of this recursive formula to the largest optimal solution.[1] Unfortunately, this optimizing formula occasionally converges to 0, and the resulting Dirac spike estimate (optimum h=0) is a degenerate estimate.[1] However, since kernel estimators are not generally robust against unsatisfactory values for the data-based smoothing parameter,[5] Tapia and Thompson's recursive formula is advantageous since it minimizes this problem. Appendix C contains the FORTRAN code programming this recursive formula for the Gaussian kernel.

Using Tapia and Thompson's formula on the ATI Auto sample produces a Dirac spike estimate (i.e., optimum h=0). But, using Tapia and Thompson's formula on the ATI Review sample suggests that the optimum h value is approximately 1.02, and this

---

[6]Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, p. 60.

19

optimum Gaussian kernel density estimate in **Figure 7** "fits" the ATI Review sample's histogram well.

Although **Figure 7** depicts this sample with five significant modes in the interval [0,50] (and thirteen significant modes for the entire range of the sample), kernel estimators cannot take into account all the theoretical and experimental conditions producing a data sample. As Parzen noted, after a reliable kernel density estimate for a sample is obtained, one should investigate the theoretical and experimental causes of the suggested modes. Parzen suggests that this inquiry may justify adjusting the value of the smoothing parameter in some instances.[7] For example, since FIST FDT&E II service times were only resolved to the nearest second, further analysis may suggest adjusting the value of the smoothing parameter. For the purposes of this report, analysis to determine the optimum h values based on theoretical and experimental conditions will not be undertaken and should not detract from the *a priori* power of kernel density estimation.

Monte Carlo simulations for the Gaussian kernel

To evaluate the performance of the Gaussian kernel and Tapia & Thompson's recursive formula, Monte Carlo simulations were conducted from the following distributions: a) $N(0,1)$, the standard normal distribution; b) $0.5\ N(-1.5,1) + 0.5\ N(1.5,1)$, a 50-50 mixture of two normal distributions; c) an F distribution with $(10,10)$ degrees of freedom; d) a Gamma distribution with $f(x) = x * \exp(-x)$ for x between 0 and ∞; and e) an Exponential distribution with parameter $\theta = 1$. In order to evaluate the performance of the Gaussian kernel and Tapia & Thompson's recursive formula, average efficiency was selected and was computed by dividing the theoretical IMSE by the mean IMSE. Results for Monte Carlo simulations for distributions a through e were reported by both Tapia & Thompson, and Scott & Factor. Analogous to their simulations,[8] twenty-five pseudo-random samples with a sample size of n were generated for each distribution. **Table 4** presents the results from these Monte Carlo simulations.

The results for the first three distributions in **Table 4** are close to the results reported by Scott & Factor, and Tapia & Thompson.[8] With the exception of the Exponential distribution, the mean optimum h values for each distribution in **Table 4** conform relatively well with the theoretical optimum h values. As the sample size, n, increased, the range of the optimum h values decreased. As anticipated, the average efficiency increased as n increased for the first five distributions and was best for Monte Carlo simulations from the $N(0,1)$ distribution. The latter result was expected since the standard normal probability density function is used for the Gaussian kernel function. Despite the reduced average efficiency of the mixed normal distribution in comparison to

---

[7] Parzen, Emanual. "Comment," *Journal of the American Statistical Association*, vol. 75, no. 369, March 1980, p. 59.

[8] Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981, pp. 11-12. and Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, pp. 67-68.

Gaussian Kernel

ATI Review

h = 1.02

Service Time (sec)

Number of Observations
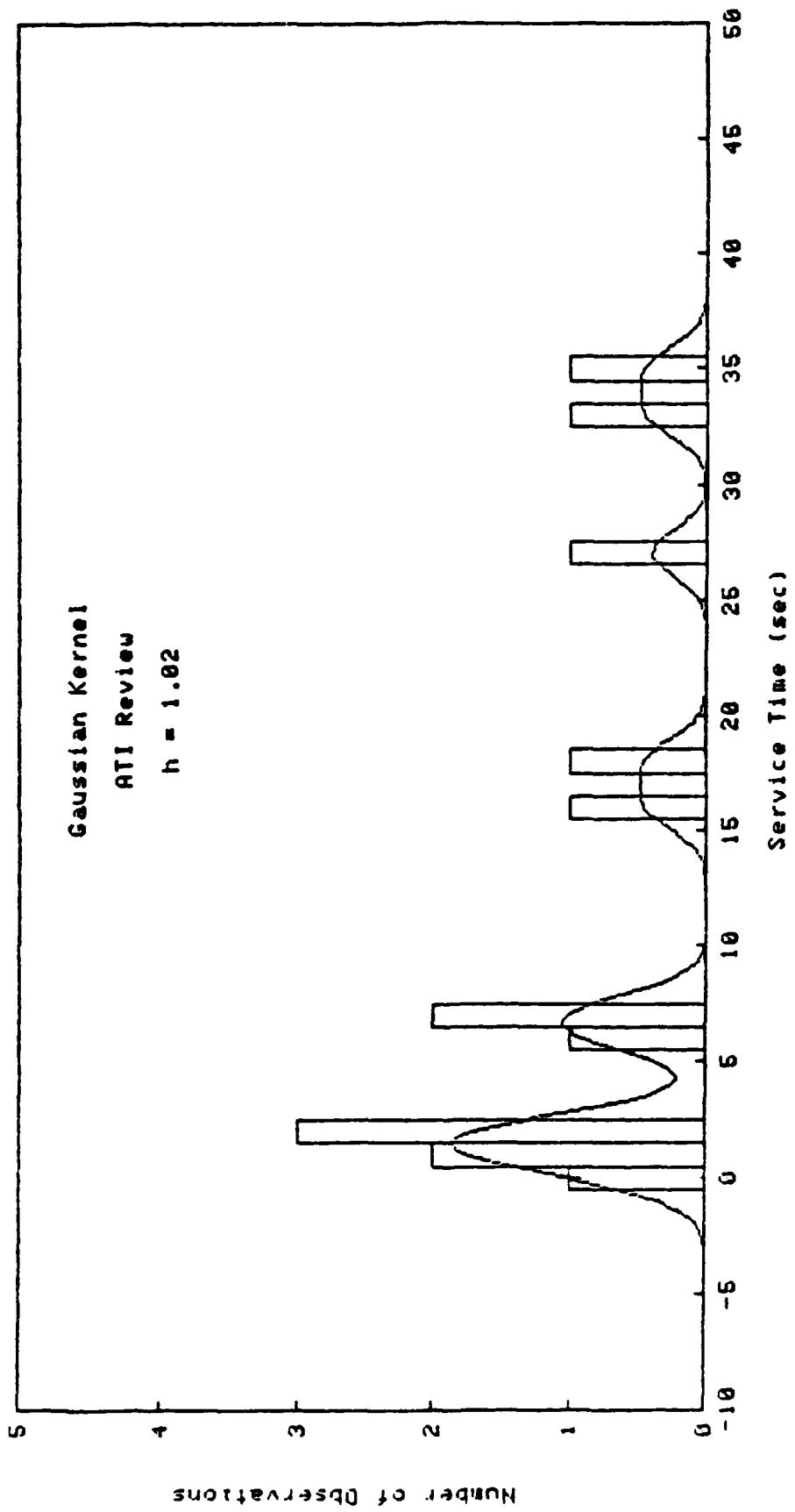
Figure 7. Optimum Application of the Gaussian Kernel to ATI Review Sample h=1.02.

TABLE 4. Monte Carlo Simulations Employing the Gaussian Kernel
(Each row represents 25 samples of size n.)

Distribution	Sample Size (n)	Degenerate Samples	Range for Optimum h	Mean Optimum h	Theoretical Optimum h	Mean IMSE	Theoretical IMSE	Average Efficiency (%)
N(0,1)	10	0	0.08 - 1.05	0.65	0.67	0.08951	0.05276	58.94
	25	0	0.18 - 0.89	0.50	0.56	0.03286	0.02535	77.15
	50	0	0.17 - 0.59	0.41	0.48	0.01773	0.01456	81.89
	100	0	0.20 - 0.49	0.35	0.42	0.00958	0.008362	87.37
0.5 N(1.5,1) + 0.5 N(-1.5,1)	10	0	0.20 - 1.87	0.94	0.79	0.09759	0.04449	45.59
	25	0	0.29 - 1.31	0.74	0.66	0.03652	0.02138	58.54
	50	0	0.13 - 0.97	0.55	0.57	0.01881	0.01228	65.28
	100	0	0.25 - 0.83	0.45	0.50	0.008971	0.007051	78.50
$F_{10,10}$	10	0	0.07 - 0.66	0.37	0.24	0.4626	0.1479	31.97
	25	1	0.04 - 0.36	0.24	0.20	0.1055	0.07107	67.36
	50	0	0.04 - 0.24	0.17	0.17	0.05181	0.04187	80.81
	100	0	0.06 - 0.22	0.15	0 15	0.02783	0.02344	84.23
Gamma	10	0	0.16 - 1.16	0.65	0.47	0.3140	0.07527	23.97
	25	0	0.07 - 0.74	0.44	0.39	0.07805	0.03616	46.33
	50	0	0.08 - 0.49	0.33	0.34	0.03324	0.02077	62.49
	100	0	0.12 - 0.45	0.28	0.30	0.01774	0.01193	67.25
Exponential $\theta = 1$	10	0	0.09-0.71	0.37	0.56	0.1119	0.06267	56.01
	25	0	0.02-0.38	0.19	0.48	0.09286	0.03014	32.46
	50	0	0.02-0.23	0.14	0.41	0 05349	0.01729	32.32
	100	0	0.02-0.16	0.09	0.35	0 03974	0 009935	25.00

the standard normal distribution, the Gaussian kernel and Tapia & Thompson's formula are reliable for this mixed normal distribution.

The F. Gamma and Exponential distributions in **Table 4** were chosen to show the performance of the Gaussian kernel and Tapia & Thompson's formula on strictly nonnegative samples. The results for the F and Gamma distributions indicate reliable performance; however, the Gaussian kernel and Tapia & Thompson's formula show reduced performance for the Exponential distribution as n increases. The density function for an Exponential distribution with parameter $\theta = 1$ is concentrated close to zero, and, therefore, it is not surprising that the symmetric Gaussian kernel has reduced average efficiency.

The Monte Carlo results in **Table 4** support Scott and Factor's statement that extensive numerical work has shown that for sample sizes as small as twenty-five, the mean optimum h values based on Tapia & Thompson's formula are on average close to the theoretical optimum h values.[9]

Criticisms of the Gaussian kernel

The Gaussian kernel is one of the popular unimodal kernel functions that is symmetric about zero, and it is now known that many symmetric unimodal kernel functions are nearly optimal.[1] Although the Gaussian kernel is efficient and reliable, **Figures 1** through **7** show that the Gaussian kernel density estimator predicts non-zero densities for service times less than zero seconds. Since negative service times are theoretically and experimentally impossible, one can conclude that the Gaussian kernel may not be the best kernel to use for a strictly non-negative sample with values "close" to or equal to zero. Therefore, the Gaussian kernel will be truncated below a value considered to be "too close" to zero in order to "create" a non-symmetric unimodal kernel function.

III. TRUNCATED GAUSSIAN KERNEL

Definition

For strictly non-negative samples, a truncated normal curve for sample values lying below 3*h seemed to be a plausible modification to the accepted Gaussian kernel. This truncated Gaussian kernel density estimator is defined as follows:

For z: $z \in \mathbf{R}, 0 \le z < \infty$

$$\hat{f}(z) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{z-x_j}{h}\right)$$

where

For $x_j < 3*h$,

$$K\left(\frac{z-x_j}{h}\right) = 0 \qquad \text{when } z < 0$$

$$K\left(\frac{z-x_j}{h}\right) = \frac{\frac{1}{\sqrt{2\pi}\,nh} e^{-1/2(\frac{z-x_i}{h})^2}}{\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\,nh} e^{-1/2(\frac{y-x_i}{h})^2} dy} \qquad \text{when } z \ge 0$$

[9] Scott, D. W., Tapia, R. A. and Thompson, J. R., "Kernel Density Estimation Revisited," *Journal of Nonlinear Analysis, Theory, Methods and Applications*, vol 1, 1977 cited by Scott, David W and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981 p. 10

For $x_j \geq 3*h$,

$$K\left(\frac{z-x_j}{h}\right) = 0 \qquad \text{when } z < 0$$

$$K\left(\frac{z-x_j}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-1/2(\frac{z-x_i}{h})^2} \text{ when } z \geq 0 \; .$$

Appendix D contains the FORTRAN code programming the truncated Gaussian kernel's estimate of a sample's underlying probability density function and cumulative distribution function.

Application of the truncated Gaussian kernel to the FIST HQ samples

Table 5 presents the truncated Gaussian kernel's critical h values for the FIST HQ samples. These values were obtained by interactively selecting h values that, in turn, were used to produce kernel density estimates.

TABLE 5. Truncated Gaussian Kernel's Critical h Values
for the FIST HQ Samples

Sample	Number of Modes in interval [0.50]	Critical h Value to two decimal places
ATI Auto	1	9.81
	2	2.64
	3	1.85
	4	1.01
	5	$1.00 < h < 1.01$
	6	0.65
	7	0.50
ATI Review	1	6.03
	2	4.01
	3	2.60
	4	1.85
	5	1.00
	6	$0.99 < h < 1.00$
	7	0.48

Figures 8 through **10** present truncated Gaussian kernel density estimates for the ATI Auto sample using h equal to 2.00, 1.00 and 0.50, respectively. The density estimate in **Figure 9** with a smoothing parameter equal to 1.00 seems to be a good
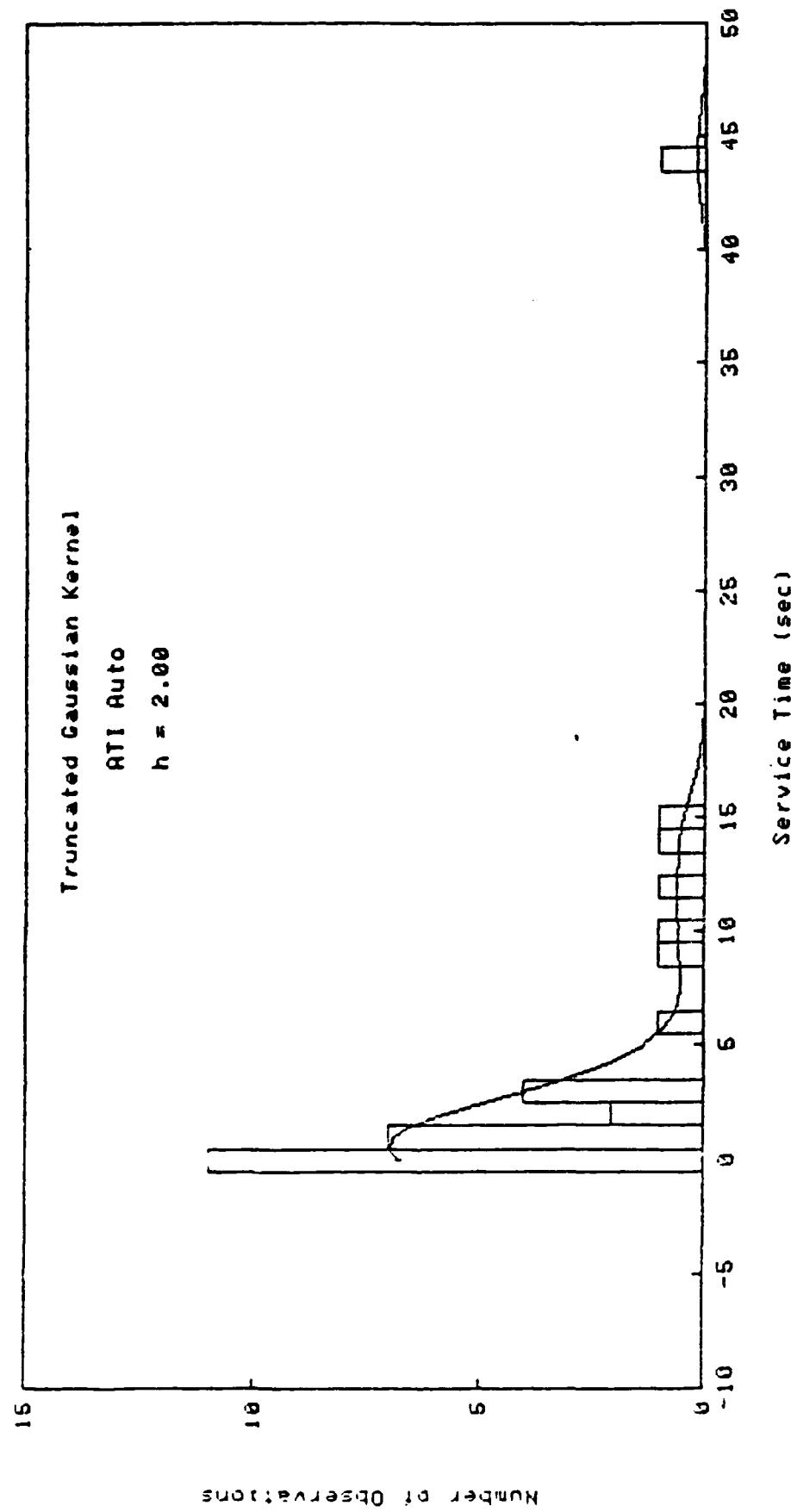
Figure 8. Application of the Truncated Gaussian Kernel with h=2.00 to ATI Auto Sample.
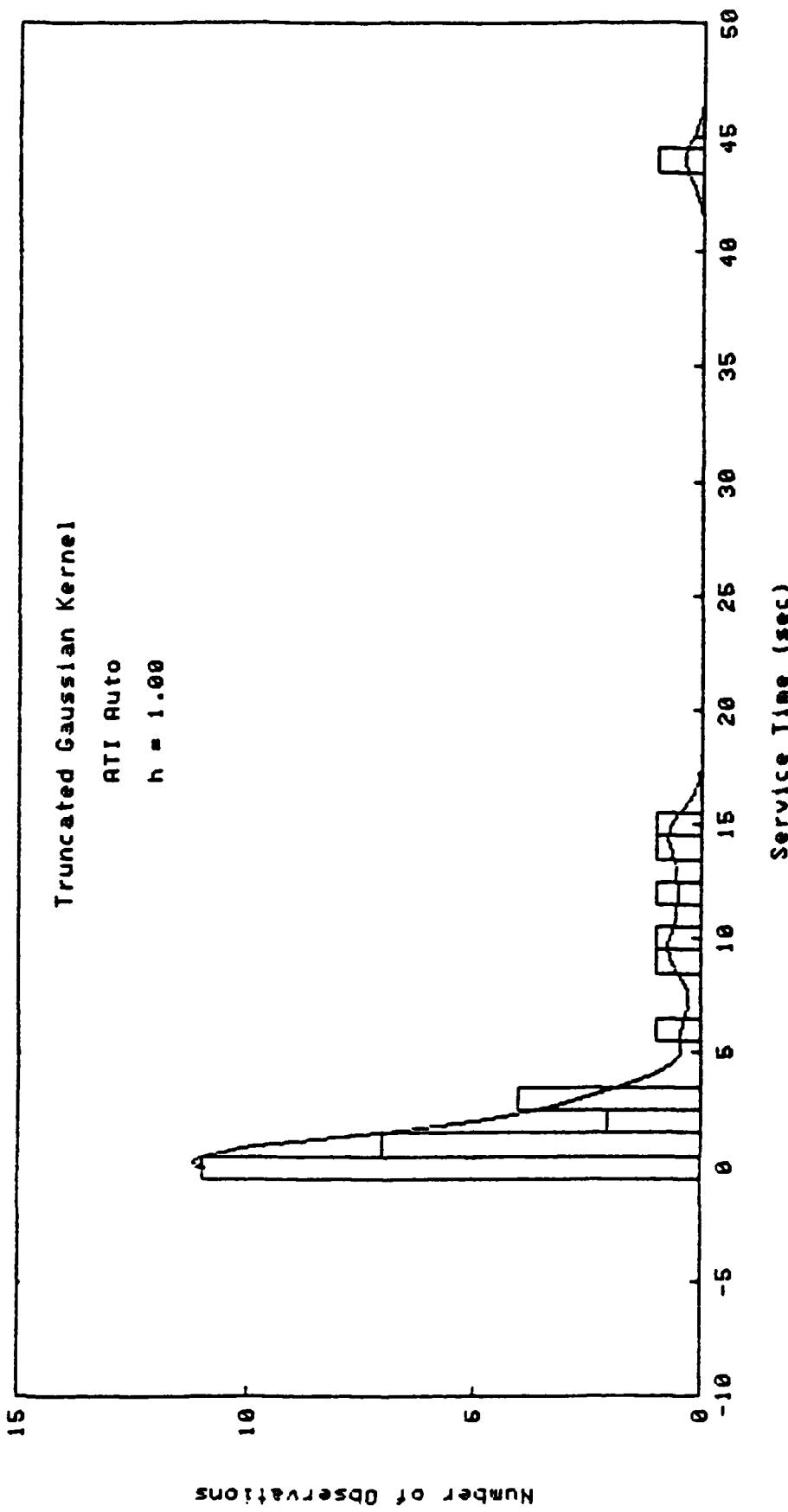
Figure 9. Application of the Truncated Gaussian Kernel with h=1.00 to ATI Auto Sample.

26
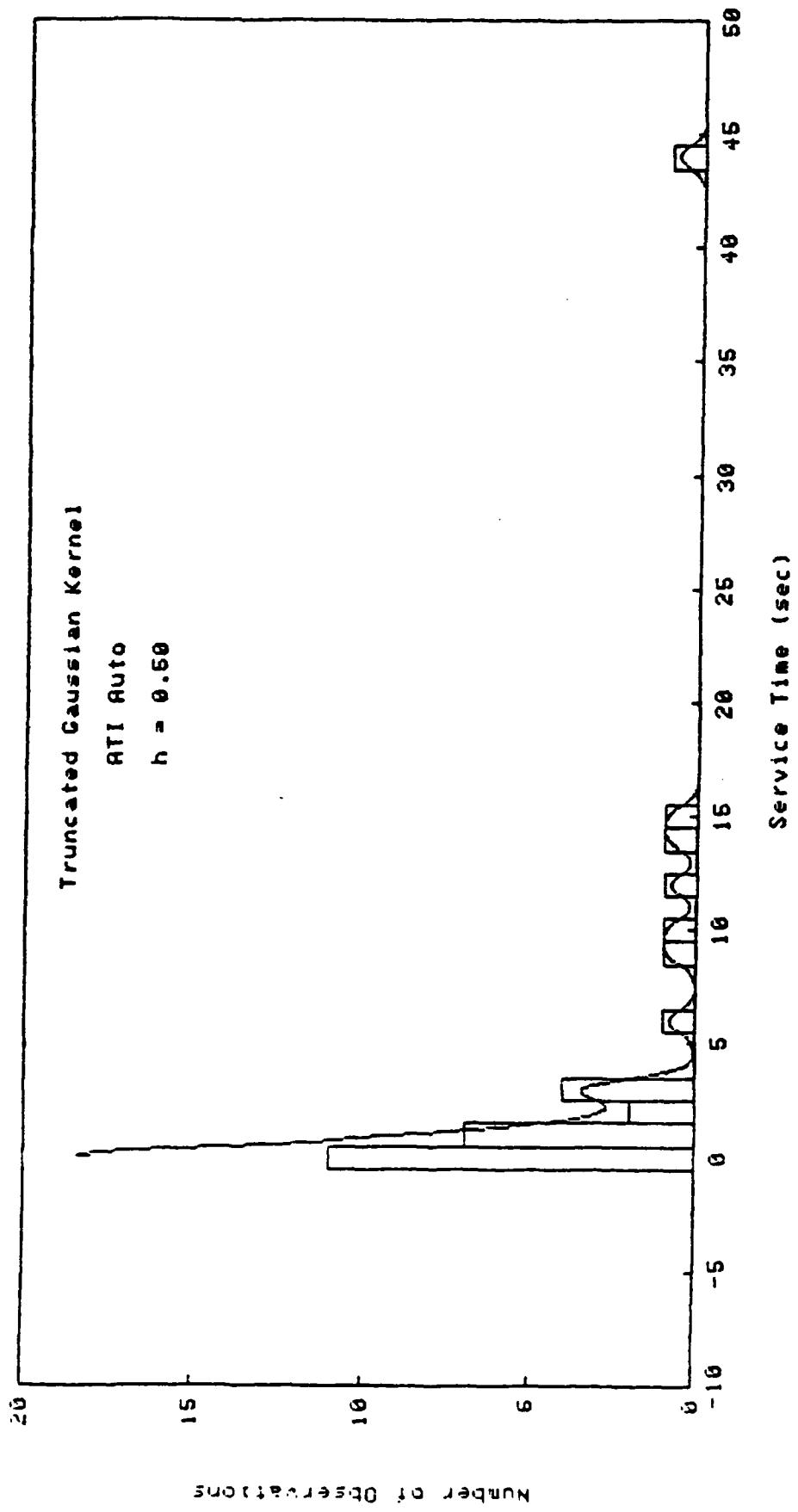
Figure 10. Application of the Truncated Gaussian Kernel with h=0.50 to ATI Auto Sample.

estimate of the underlying probability density function and predicts the same modes as the Gaussian kernel density estimate in **Figure 2**. The only main difference between these two density estimates occurs in the region below zero seconds.

Figures 11 through **13** present truncated Gaussian kernel density estimates for the ATI Review sample with h equal to 5.00, 3.00 and 1.00, respectively. The density estimate in **Figure 13** with the smoothing parameter equal to 1.00 seems to be a good density estimate.

Appendix E contains the FORTRAN code programming Tapia and Thompson's optimizing h formula for the truncated Gaussian kernel. Using this programming code on the ATI Auto sample suggested that the optimum h value is approximately 0.62. **Figure 14** compares the truncated Gaussian kernel estimate for the ATI Auto sample using this h value and the histogram for this sample. Using Tapia and Thompson's optimizing h formula on the ATI Review sample produced a Dirac spike estimate when the sample range was used for the initial h value; however, using 10.00, 5.00 and 1.00 as the initial h values each produced optimizing h values approximately equal to 1.27. Although Tapia and Thompson's formula converges to 0 with an initial h value equal to the sample's range, other initial h values suggest that a good h value produces a density estimate with 5 modes in the interval [0.50], and substantiates the earlier claim that the density estimate in **Figure 13** with a smoothing parameter equal to 1.00 is a reasonable density estimate for the ATI Review sample.

Monte Carlo simulations for the truncated Gaussian kernel

To evaluate the performance of the truncated Gaussian kernel and Tapia & Thompson's recursive formula, Monte Carlo simulations were conducted from the same distributions as the Monte Carlo simulations for the Gaussian kernel. Analogous to the Gaussian kernel simulations(**Table 4**), twenty-five pseudo-random samples, each with a sample size of n, were generated from these distributions and the h values associated with these samples were calculated. **Table 6** presents the results from these Monte Carlo simulations.

Table 6 reflects the difficulties that were associated with measuring the average efficiency of these Monte Carlo simulations. For instance, although the value for α (K) of Tapia & Thompson's formula can be evaluated for the truncated Gaussian kernel, its two different values, $\frac{1}{\sqrt{\pi}}$ for $x_i < 3*h$ and $z \geq 0$, and $\frac{1}{2\sqrt{\pi}}$ for $x_i \geq 3*h$ and $z \geq 0$, interfere with the evaluation of the the theoretical optimum h values, IMSE [f̂] values and subsequently, the average efficiency values. Thus, **Table 6** presents the theoretical ranges for the optimum h value and IMSE[f̂]. The lower and upper limits of these ranges were computed by using the two different α(K) values. The h value associated with each sample of size n, however, was computed by using the α(K) value that was correct for each sample value x_i.
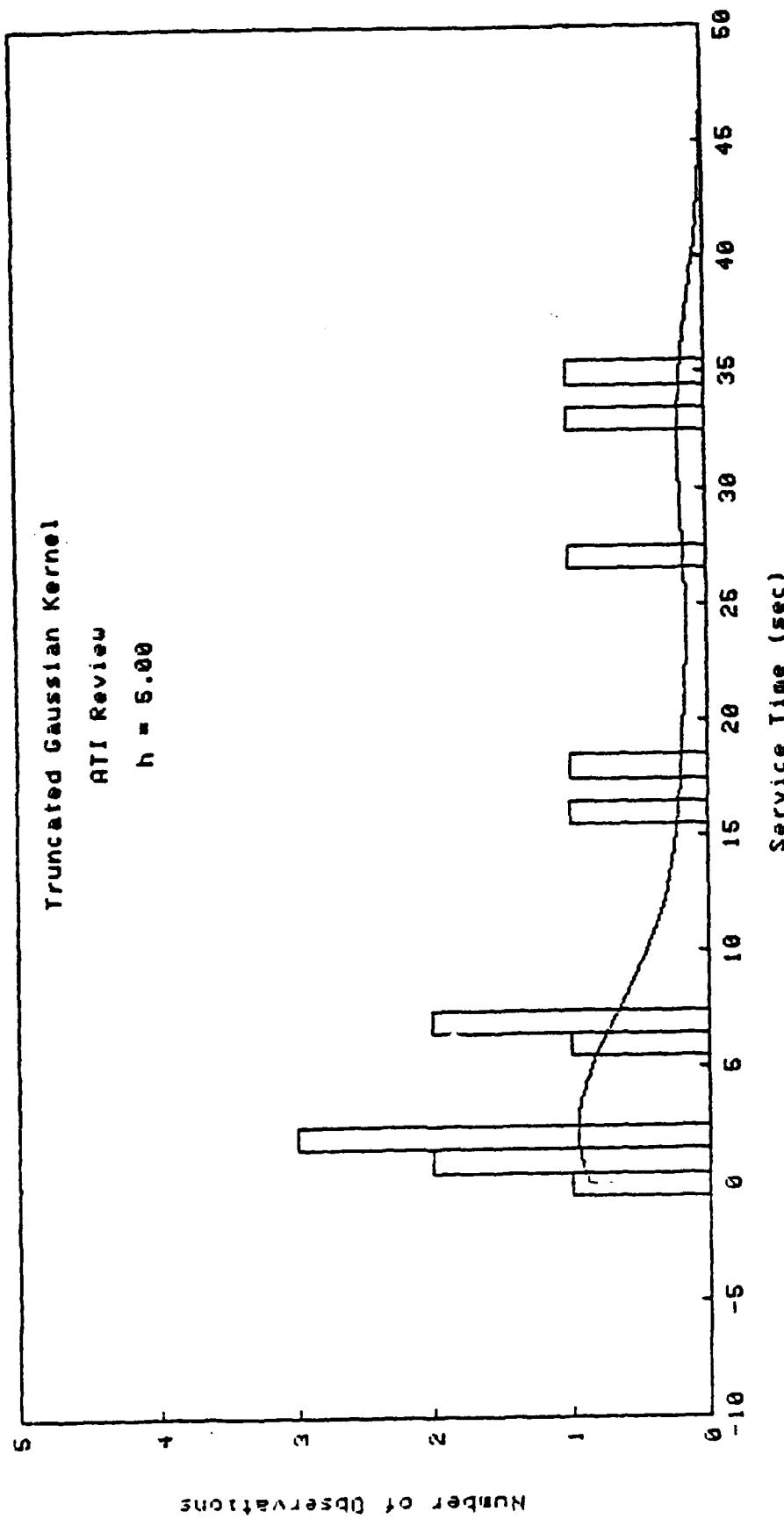
Figure 11. Application f the Truncated Gaussian Kernel with h=5.00 to ATI Review Sample.
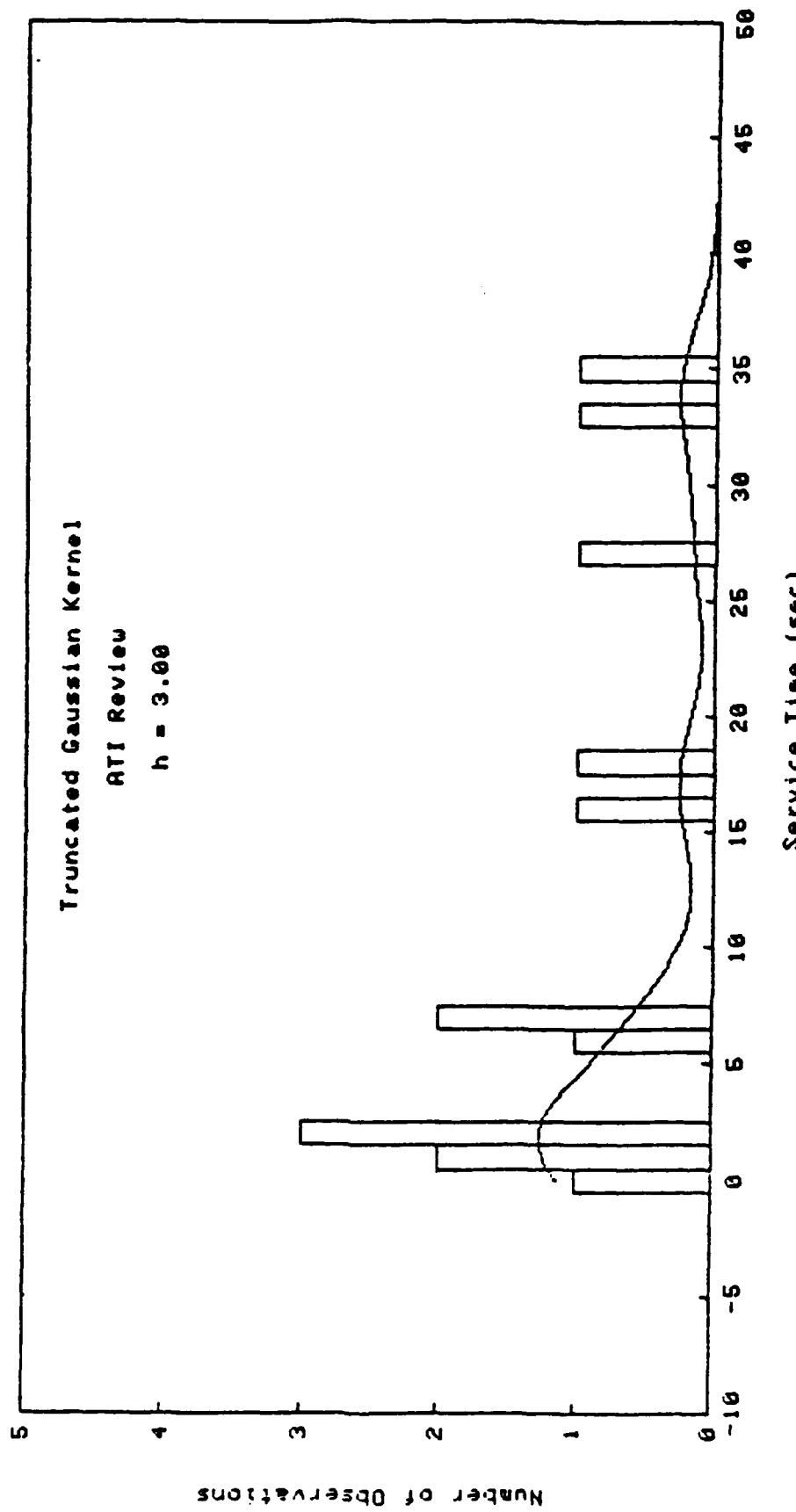
Figure 12. Application of the Truncated Gaussian Kernel with h=3.00 to ATI Review Sample.

30
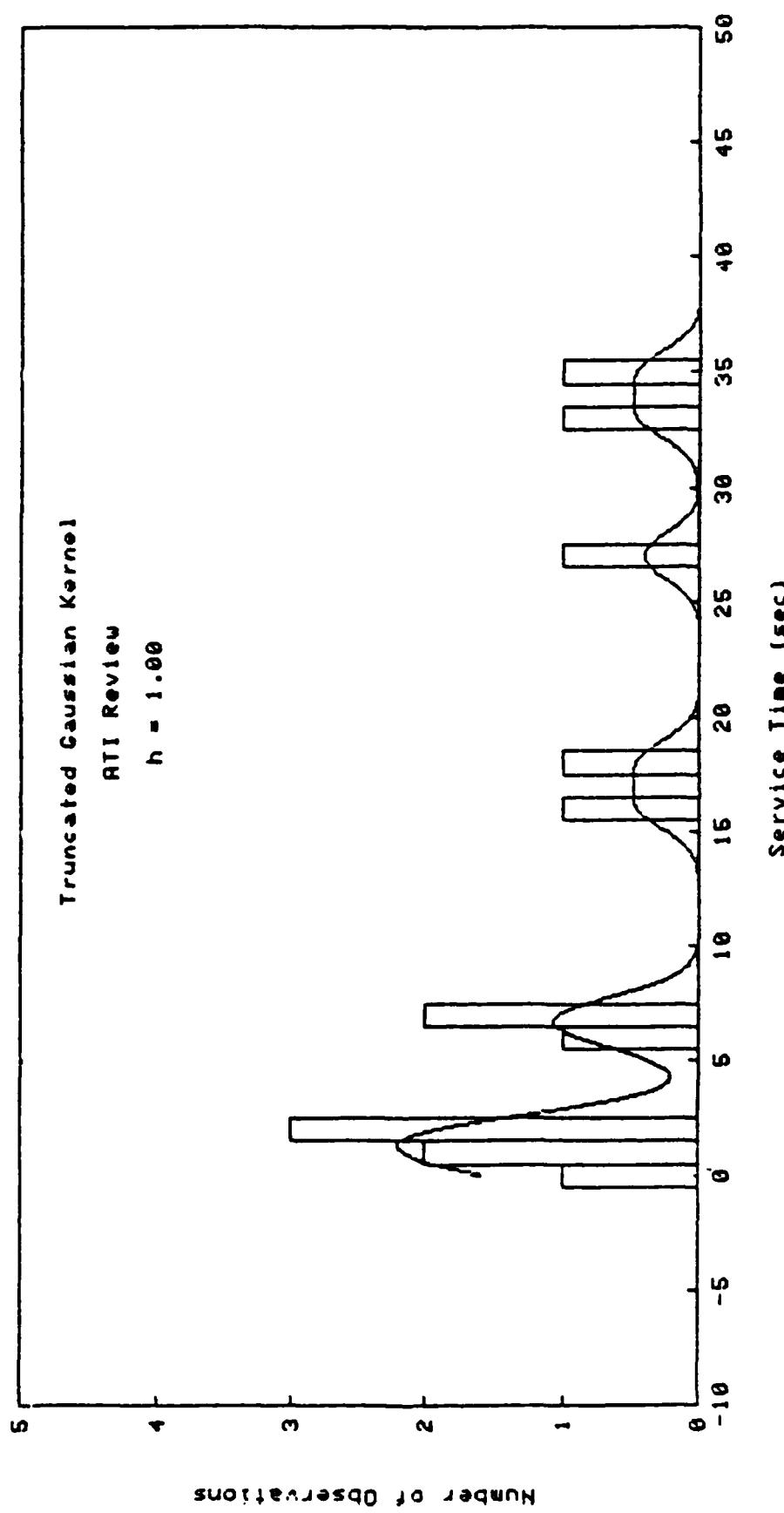
Figure 13. Application of the Truncated Gaussian Kernel with h=1.00 to ATI Review Sample.

31

Truncated Gaussian Kernel

ATI Auto

h = 0.62

Service Time (sec)

Number of Observations

20

16

10

5

0

-10   -5   0   5   10   15   20   25   30   35   40   45   50

Figure 14. Optimum Application of the Truncated Gaussian Kernel to ATI Auto Sample h=0.62.

32
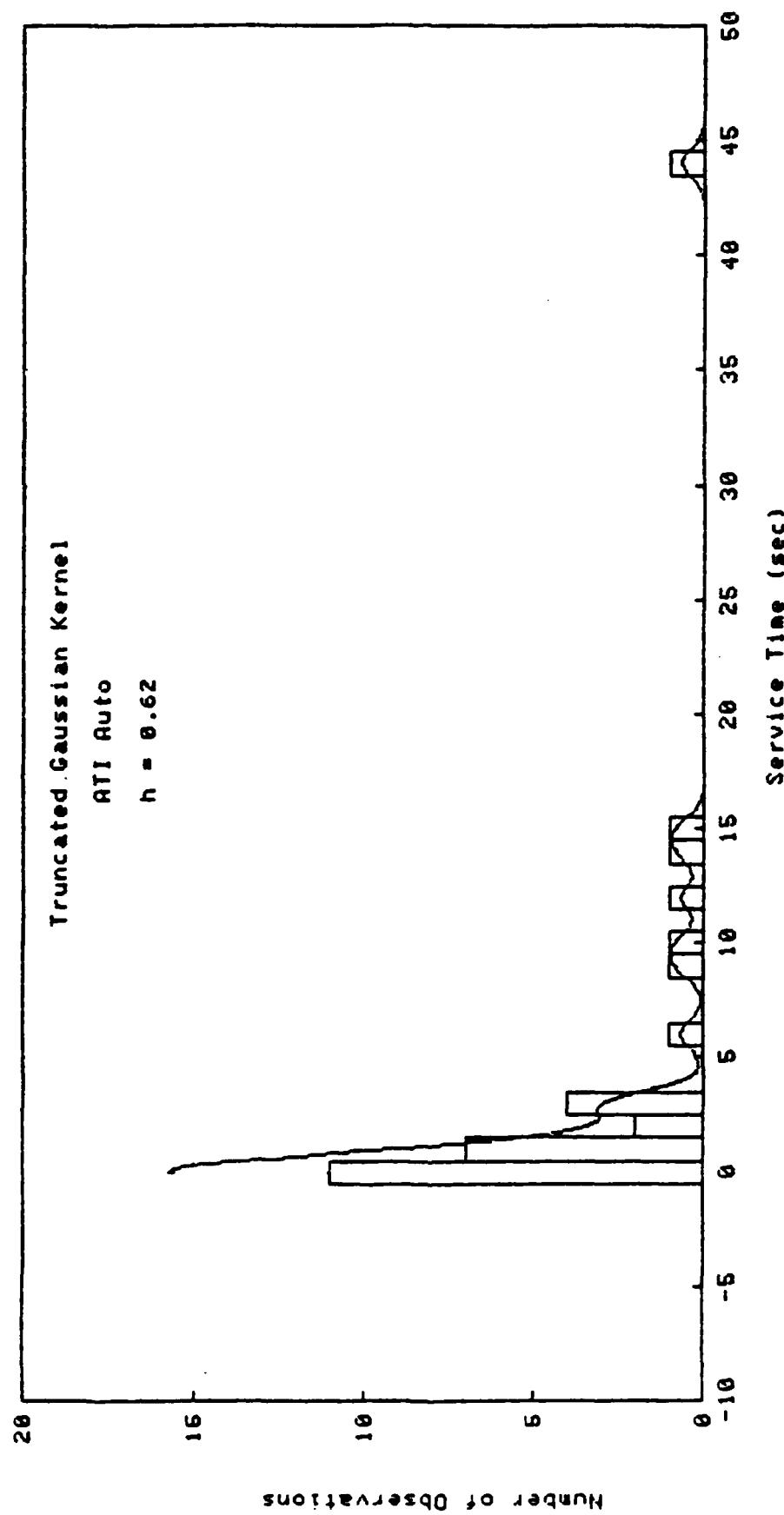
TABLE 6. Monte Carlo Simulations Employing the Truncated Gaussian Kernel
(Each row represents 25 samples of size n.)

Distribution	Sample Size (n)	Degenerate Samples	Range for Optimum h	Mean Optimum h	Range for Theoretical Optimum h	Range for Mean IMSE	Range for Theoretical IMSE	Efficiency Range (%)
$N(5,1)$	10	0	0.08-1.05	0.65	0.67-0.77	0.9102-0.1026	0.04593-0.05523	44.77-60.68
	25	0	0.09-0.89	0.49	0.56-0.63	0.03544-0.03883	0.02208-0.02624	56.83-74.04
	50	0	0.15-0.63	0.41	0.48-0.56	0.01771-0.01902	0.01268-0.01528	66.67-86.28
0.5 $N(5,1)$ + 0.5 $N(8,1)$	10	0	0.14-1.60	0.92	0.79-0.91	0.09017-0.3805	0.03873-0.04646	10.18-51.53
	25	0	0.21-1.21	0.73	0.66-0.76	0.03614-.1429	0.01861-0.02237	13.02-61.90
	50	0	0.13-0.96	0.55	0.57-0.66	0.01903-0.05470	0.01069-0.01283	19.54-67.42
$F_{10,10}$	10	0	0.44-0.78	0.58	0.24-0.27	1.229-2.409	0.1288-0.1531	5.35-12.46
	25	0	0.17-0.62	0.45	0.20-0.23	0.4985-0.9699	0.06187-0.07467	6.38-14.98
	50	0	0.08-0.49	0.41	0.17-0.20	0.2770-0.5398	0.3554-0.04288	6.58-15.47
Gamma	10	0	0.16-0.99	0.72	0.47-0.54	0.1022-0.1583	0.06553-0.07881	41.40-77.11
	25	0	0.28-0.78	0.61	0.39-0.45	0.04693-0.07418	0.03148-0.03789	42.44-80.74
	50	0	0.17-0.60	0.49	0.34-0.39	0.02477-0.03646	0.01808-0.02170	49.59-87.61
Exponential $\theta=1$	10	0	0.31-0.85	0.61	0.56-0.65	0.05905-0.06946	0.05456-0.06571	78.55-100.00
	25	0	0.30-0.64	0.48	0.47-0.54	0.02800-0.03191	0.02621-0.03153	82.14-100.00
	50	0	0.31-0.45	0.41	0.41-0.47	0.01505-0.01810	0.01505-0.01810	85.08-100.00

Although the efficiency range in **Table 6** cannot be directly compared to the average efficiency in **Table 4**, it is helpful in drawing some tentative conclusions. First, preliminary comparisons between these tables suggest that as expected, the truncated Gaussian kernel may not be quite as efficient as the Gaussian kernel for normal or mixed normal distributions. Second, the source of the alarmingly low efficiency range of the F distribution with (10.10) degrees of freedom in **Table 6** has not been identified and seems counter intuitive. Third, applying the truncated Gaussian kernel to Monte Carlo simulations from the Gamma and Exponential distributions would appear to result in efficiency ranges which equal or surpass the average efficiency of the Gaussian kernel's simulations. Additional Monte Carlo simulations would be helpful in determining the performance of Tapia & Thompson's for the truncated Gaussian kernel in comparison to the Gaussian kernel. More beneficial, however, would be a more sophisticated procedure for computing the average efficiency.

Criticisms of the truncated Gaussian kernel

From the truncated Gaussian kernel figures for the FIST HQ samples, it is apparent that this kernel density estimator provides a probability density function estimate with zero density for all negative values. This property conforms perfectly with the theoretical and experimental service time expectations underlying the FIST HQ samples. Preliminary Monte Carlo simulations with skewed non-negative distributions suggested that the use of Tapia & Thompson's formula with the truncated Gaussian kernel was more efficient than the Gaussian kernel. Thus, this supports the use of the truncated Gaussian kernel for the FIST HQ samples. To further examine these FIST HQ samples with another non-symmetric, unimodal kernel function, it was decided that a Lognormal kernel, which can only be applied to strictly positive samples, should also be developed.

IV. MODE CENTERING LOGNORMAL KERNEL

Definition

A Lognormal kernel density estimator concentrating its kernel estimate around the sample's mode is defined as follows:

For z: $z \in \mathbf{R}$, $0 < z < \infty$

$$\hat{f}(z) = \frac{1}{\sqrt{2\pi}hz} \sum_{j=1}^{n} e^{\left[\frac{-1}{2}\left\{\frac{\log_e(z)-h^2-\log_e(x_j)}{h}\right\}^2\right]}$$

Unlike the normal distribution which is symmetric, the Lognormal distribution is skewed and can only be evaluated for strictly positive values. Appendix F contains the FORTRAN code programming the mode centering Lognormal kernel's estimate of a sample's underlying probability density function and cumulative distribution function.

Application of the mode centering Lognormal kernel to the FIST HQ samples

Due to the definition of the Lognormal distribution, it was necessary to modify the zero second FIST HQ service times. Although FIST HQ service times cannot be theoretically less than or equal to zero seconds, FIST HQ service times were only resolved to the nearest second. Thus, a zero second service time is actually a service time of less than one second. Therefore, the interpretation of zero second service times as half second service times seems plausible.

Table 7 presents the mode centering Lognormal kernel's critical h values for the FIST HQ samples. Comparing these values with the critical h values for the Gaussian and truncated Gaussian kernels in **Tables 3** and **5**, one can see the higher sensitivity of the smoothing parameter value for the mode centering Lognormal kernel. These values

were obtained by interactively selecting h values that, in turn, were used to produce kernel density estimates.

TABLE 7. Mode Centering Lognormal Kernel's Critical h Values
for the FIST HQ Samples

Sample	Number of Modes in interval [0,50]	Critical h Value to two decimal places
ATI Auto	1	0.38
	2	0.35
	3	0.30
	4	0.24
	5	0.20
	6	0.18
	7	0.14
ATI Review	1	0.42
	2	0.36
	3	0.34
	4	0.30
	5	0.25
	6	0.10
	7	0.07

Figures 15 through **17** present kernel density estimates for the ATI Auto sample with h equal to 0.50, 0.25 and 0.22, respectively. These figures show that the application of the mode centering Lognormal kernel to this sample results in a very peaked density estimate around a half of a second and attributes an unnoticeable density estimate to the tail service time observation at forty-four seconds.

Figures 18 through **20** present kernel density estimates for the ATI Review sample with h equal to 0.50, 0.25 and 0.20, respectively. Comparing the mode centering Lognormal kernel estimate in **Figure 19** with the truncated Gaussian kernel estimate in **Figure 13**, one can see an apparent difference between these two density estimates. For instance, the truncated Gaussian kernel attributes a single mode to the service time observations between 0 and 5 seconds; whereas, the mode centering Lognormal kernel attributes three modes to these service time observations. Also, the truncated Gaussian kernel attributes three significant modes to the service time observations beyond 15 seconds; whereas, the mode centering Lognormal kernel "smears" these observations together.

Criticisms of the mode centering Lognormal kernel

Applying the mode centering Lognormal kernel to the FIST HQ samples resulted in density estimates that minimized the effect of tail observations. In addition, the modal nature of the density estimates tend to emphasize those observations closest to zero.
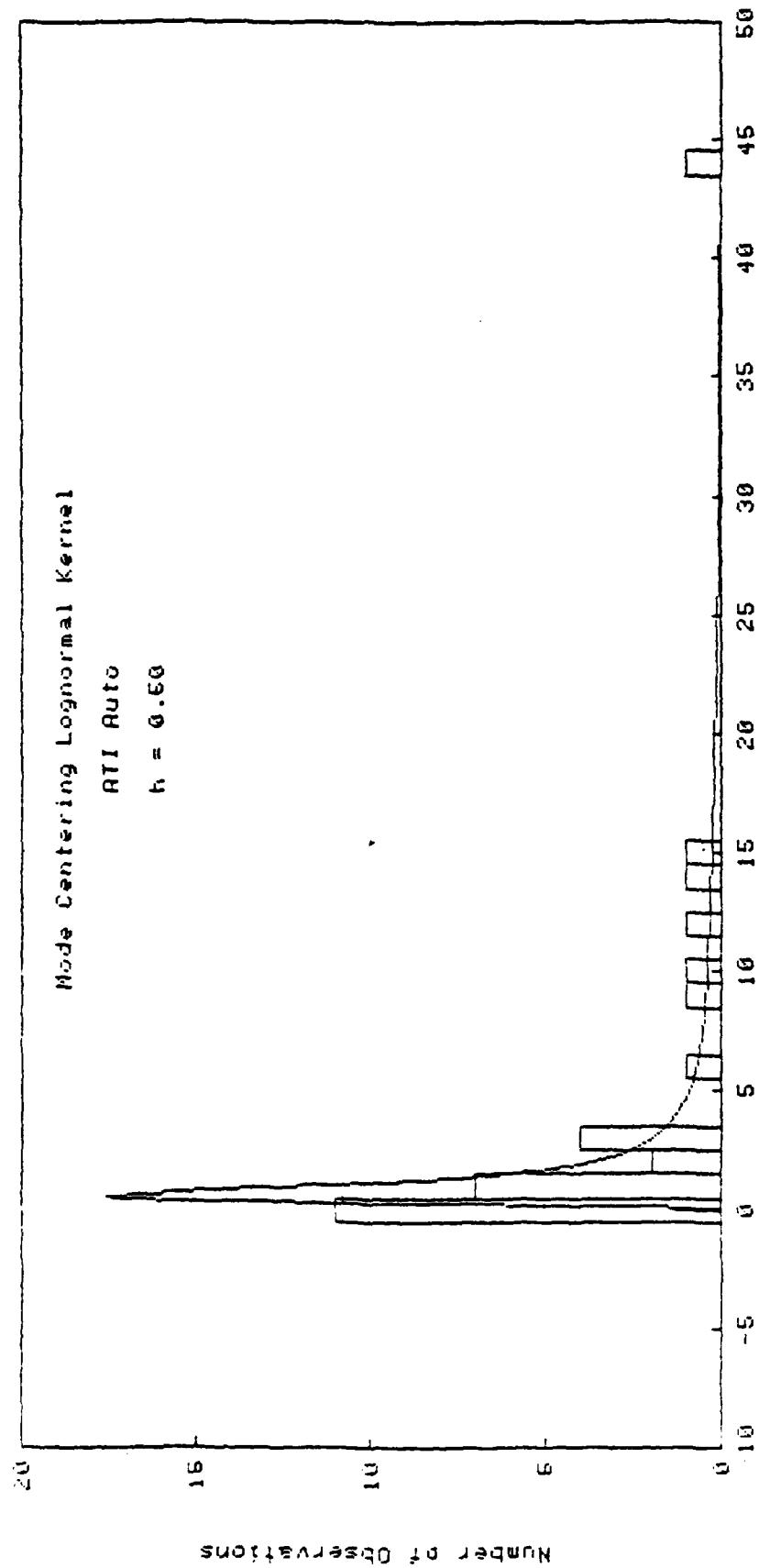
Figure 15. Application of the Mode Centering Lognormal Kernel with h= 0.50 to ATI Auto Sample.
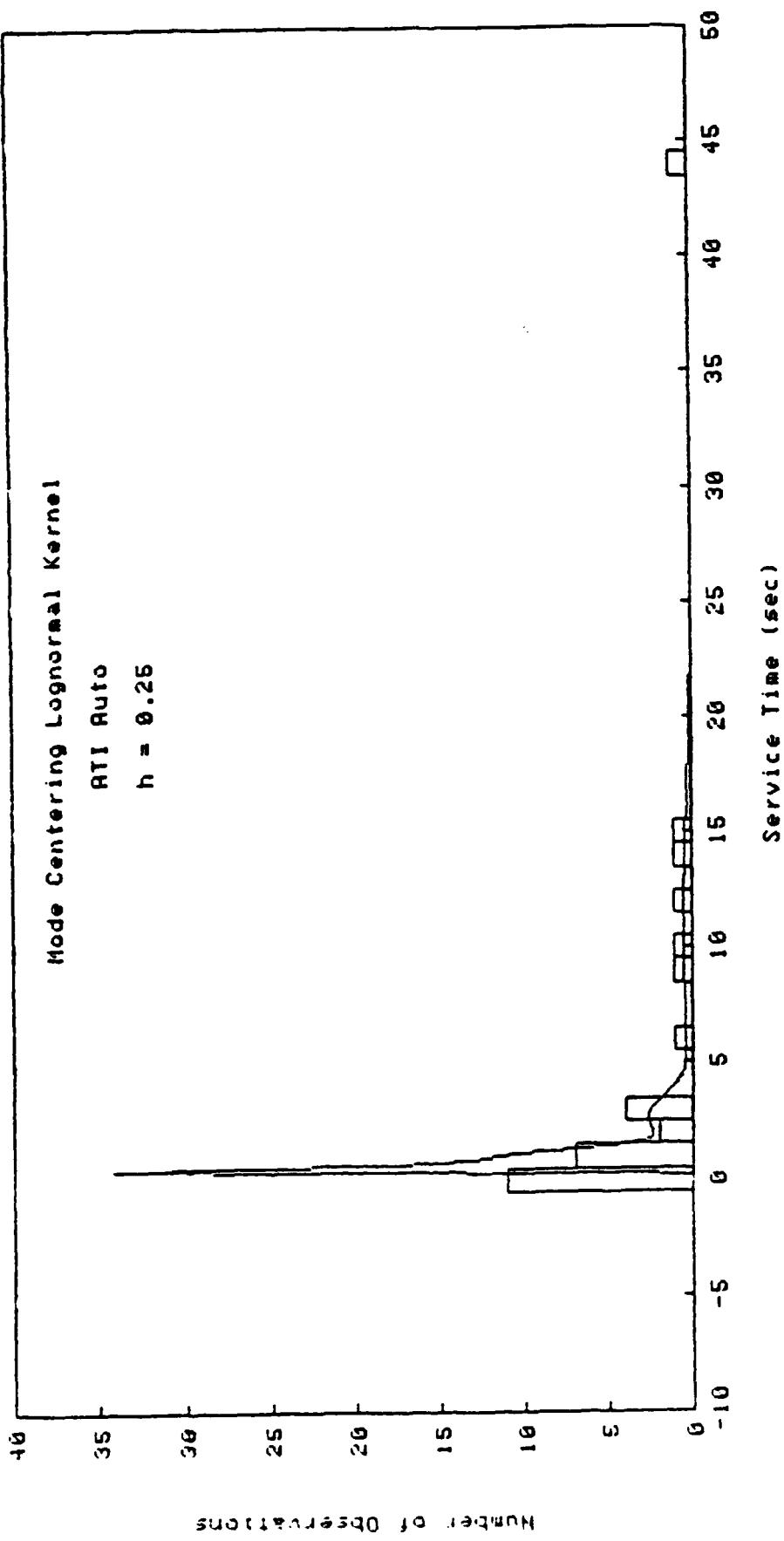
Figure 16. Application of the Mode Centering Lognormal Kernel with h=0.25 to ATI Auto Sample.
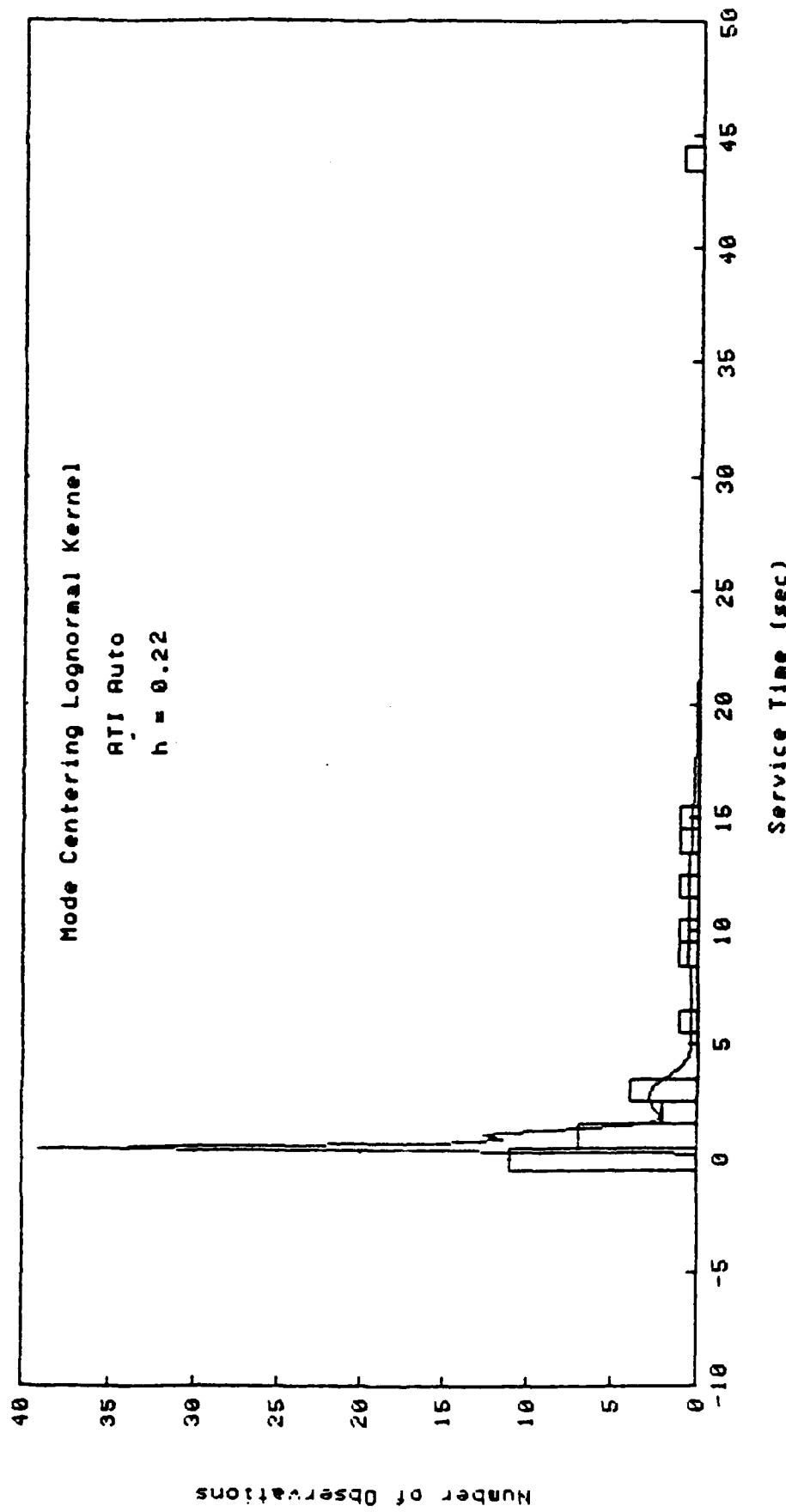
Figure 17. Application of the Mode Centering Lognormal Kernel with h=0.22 to ATI Auto Sample.
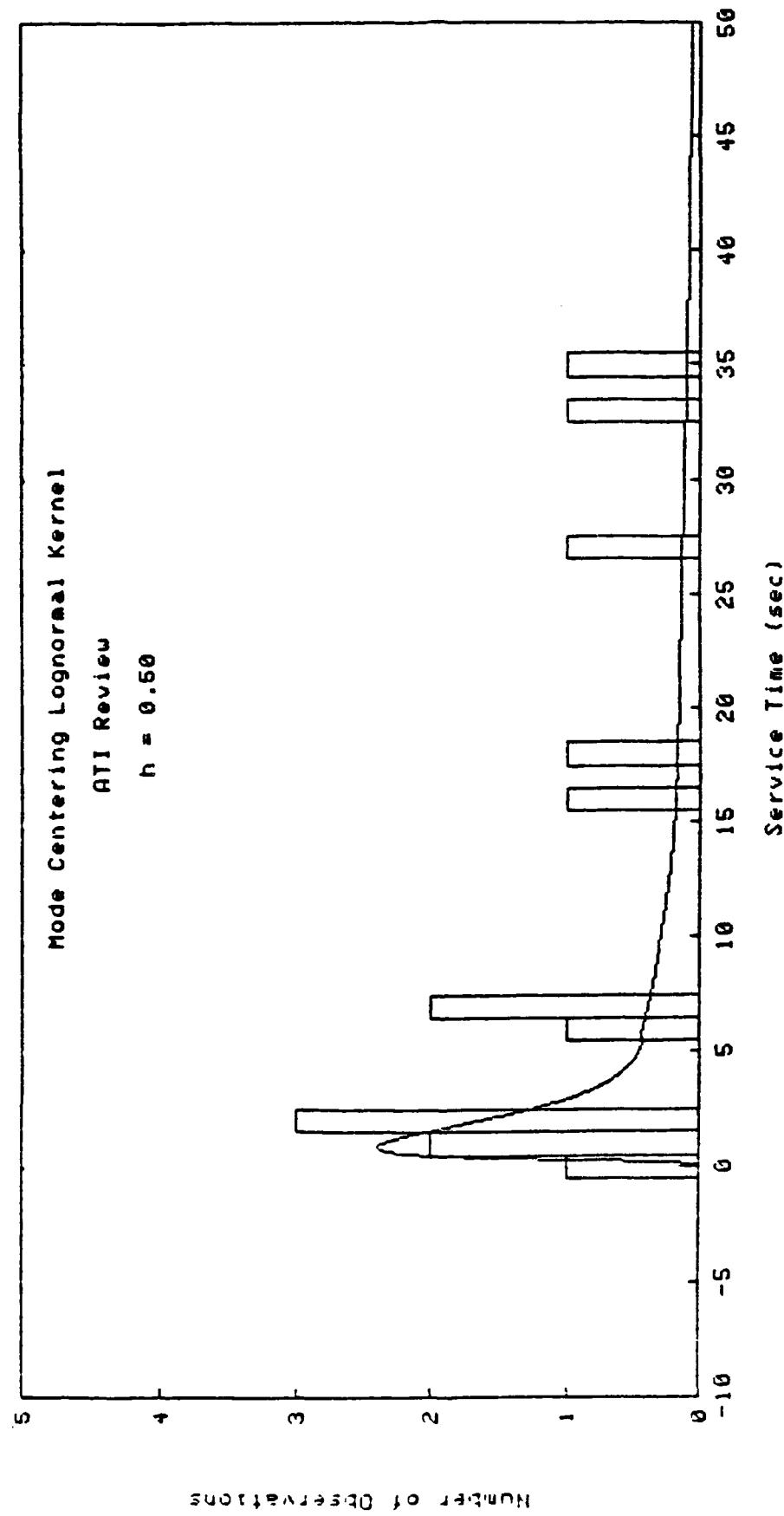
Figure 18. Application of the Mode Centering Lognormal Kernel with h=0.50 to ATI Review Sample.
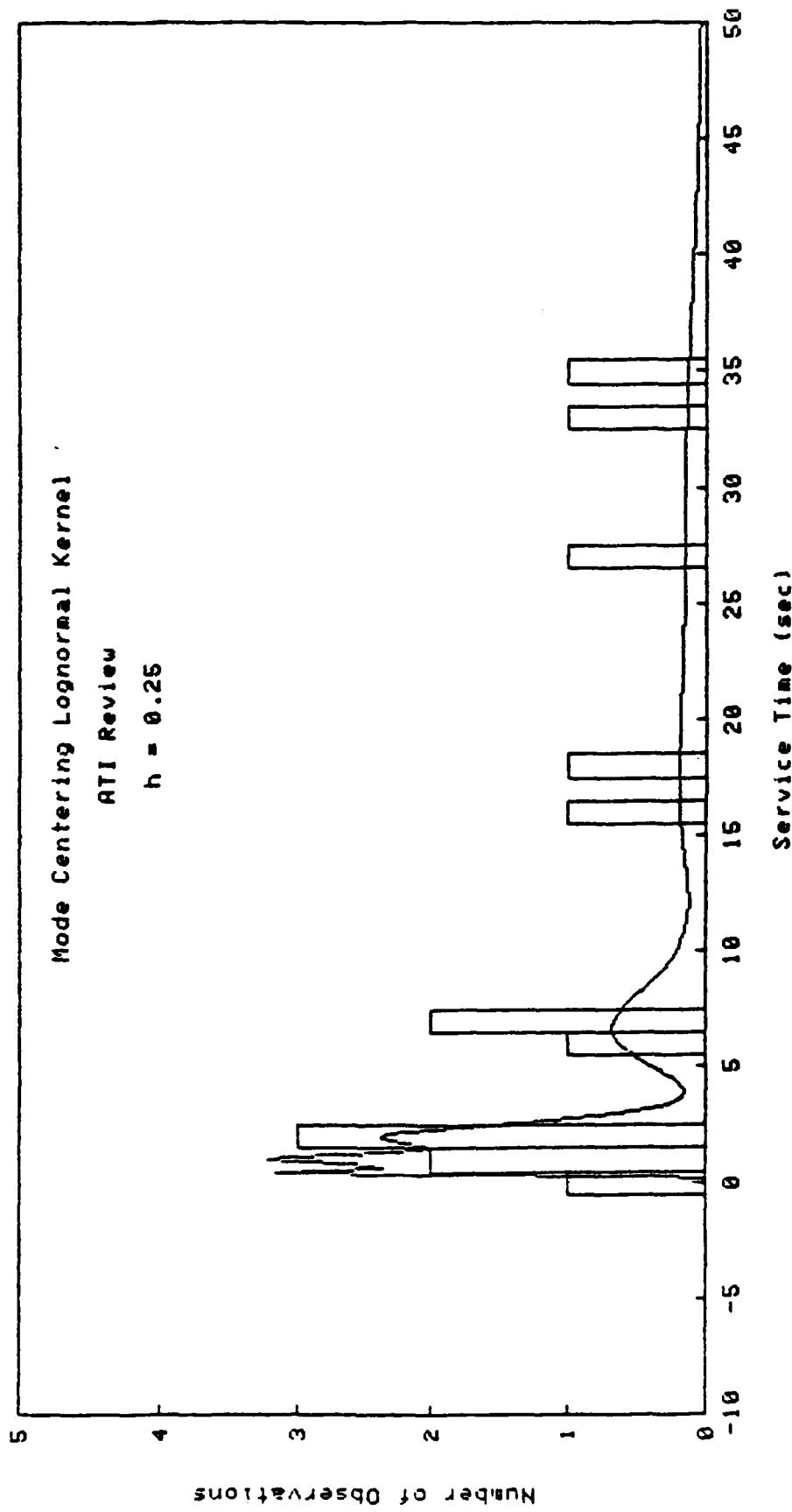
Figure 19. Application of the Mode Centering Lognormal Kernel with h=0.25 to ATI Review Sample.
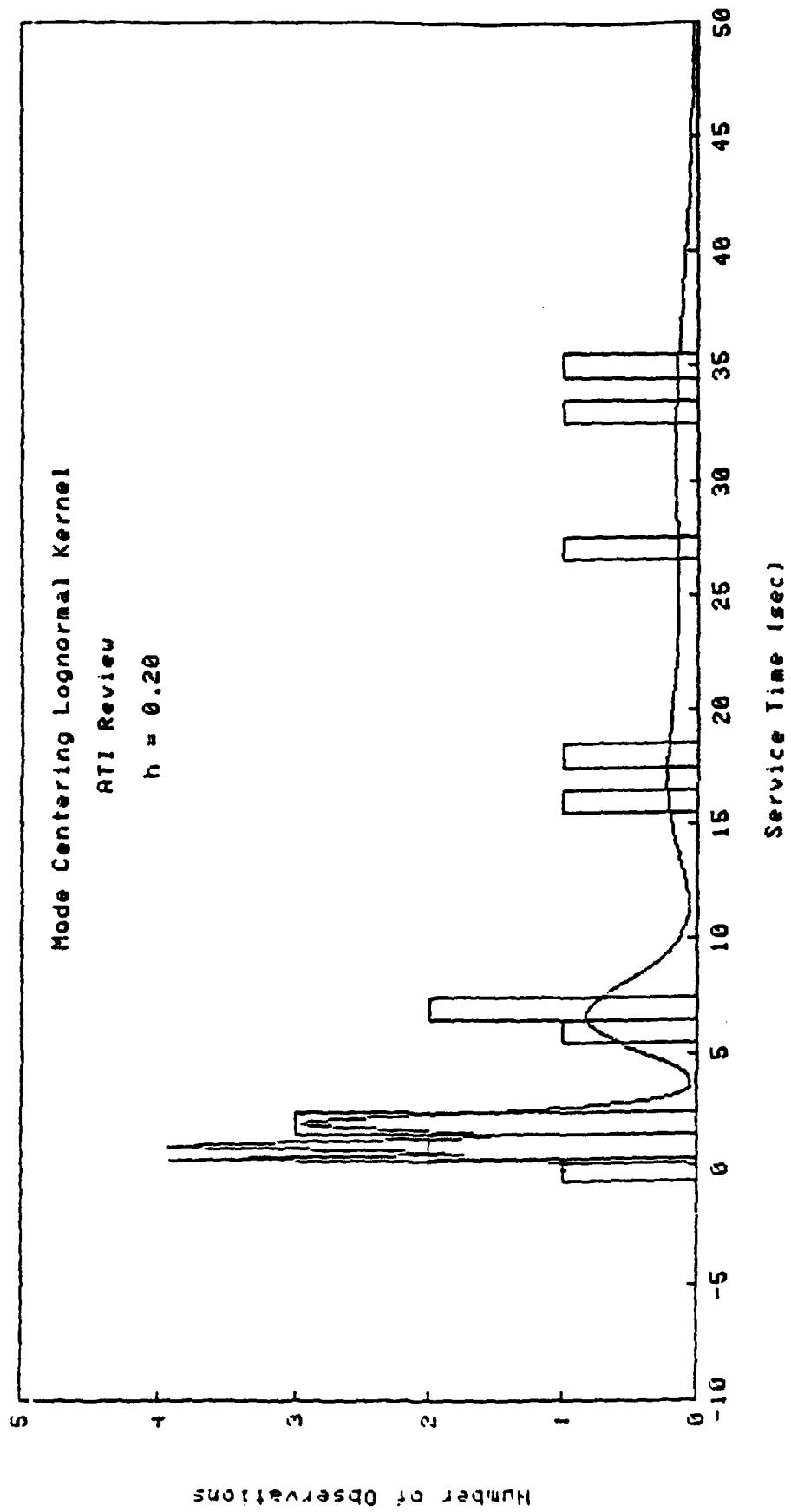
Figure 20. Application of the Mode Centering Lognormal Kernel with h=0.20 to ATI Review Sample.

Thus, the estimated underlying probability density functions associated with the mode centering Lognormal kernel differ markedly from those obtained from the Gaussian and truncated Gaussian kernels. The high sensitivity of the smoothing parameter for the FIST HQ samples to small changes could hinder the selection of an optimum smoothing parameter value, and it was decided that it would not be worthwhile to apply Tapia & Thompson's formula and to conduct Monte Carlo simulations. Instead, the median centering Lognormal kernel was developed.

V. MEDIAN CENTERING LOGNORMAL KERNEL

Definition

A Lognormal kernel density estimator concentrating its kernel estimate around the sample's median is defined as follows:

For z: $z \in \mathbf{R}, 0 < z < \infty$

$$\hat{f}(z) = \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi y}} e^{\left[\frac{-1}{2y}\log_e(z/x_j)^2\right]}$$

where

$$y = \log_e\left[\frac{1}{2}\left(1 + \sqrt{1+(2h/x_j)^2}\right)\right]$$

Appendix G contains the FORTRAN code programming the median centering Lognormal kernel's estimate of a sample's underlying probability density function and cumulative distribution function.

Application of the median centering Lognormal kernel to the FIST HQ samples

Table 8 presents the median centering Lognormal kernel's critical h values for the FIST HQ samples. These values were obtained by interactively selecting h values that, in turn, were used to produce kernel density estimates.

Figures 21 through 23 present kernel density estimates for the ATI Auto sample with h equal to 2.00, 0.95, and 0.50, respectively. Comparison of the median centering Lognormal density estimate in Figure 23 to the truncated Gaussian density estimate in Figure 14 shows that these two kernels produce similar results considering the interpretation of zero second service times to half second service times. Therefore, one can conclude that the ATI Auto sample's optimum h value for the median centering Lognormal kernel lies around 0.50.
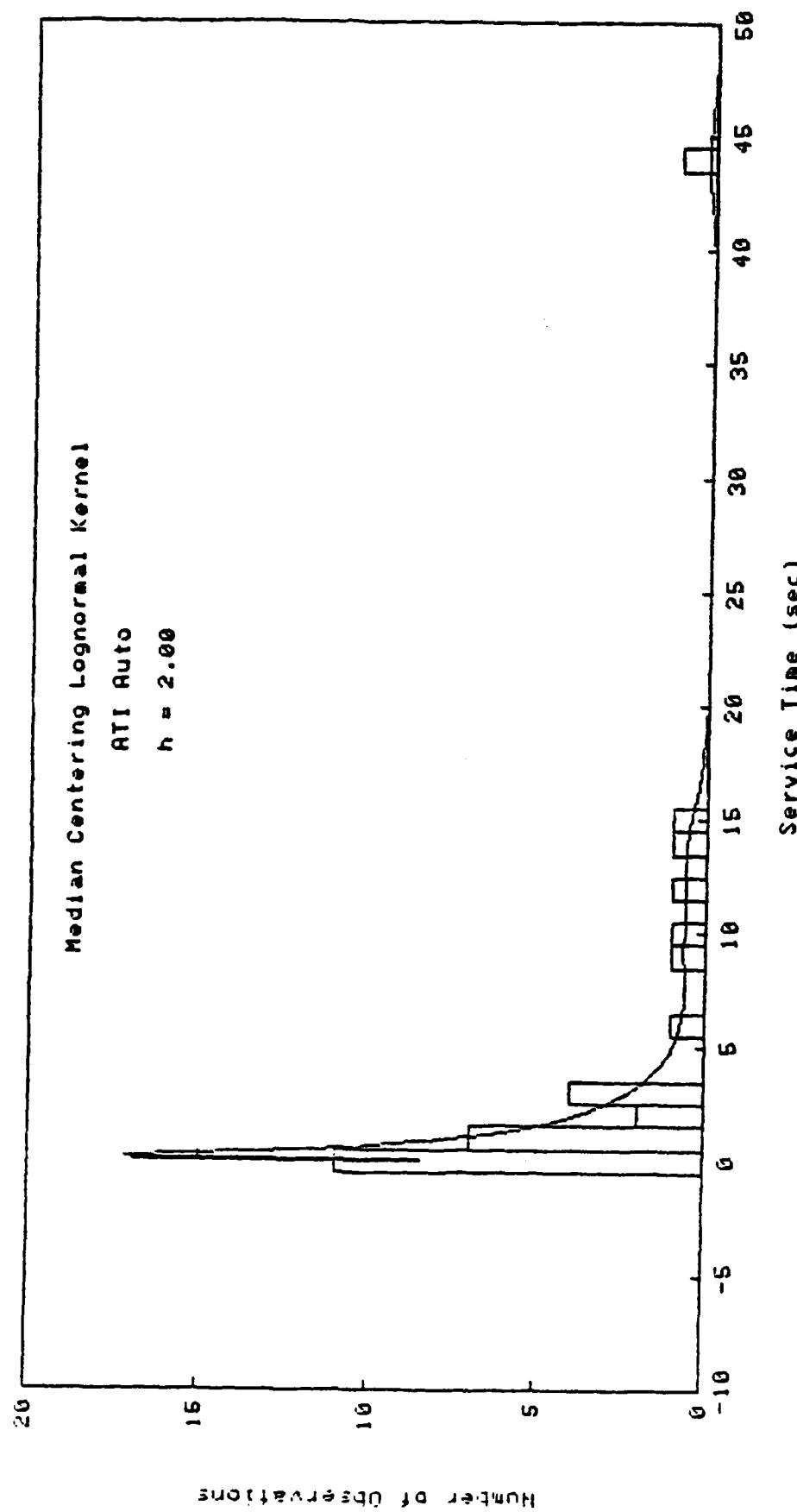
Figure 21. Application of the Median Centering lognormal Kernel with h=2.00 to ATI Auto Sample.

43
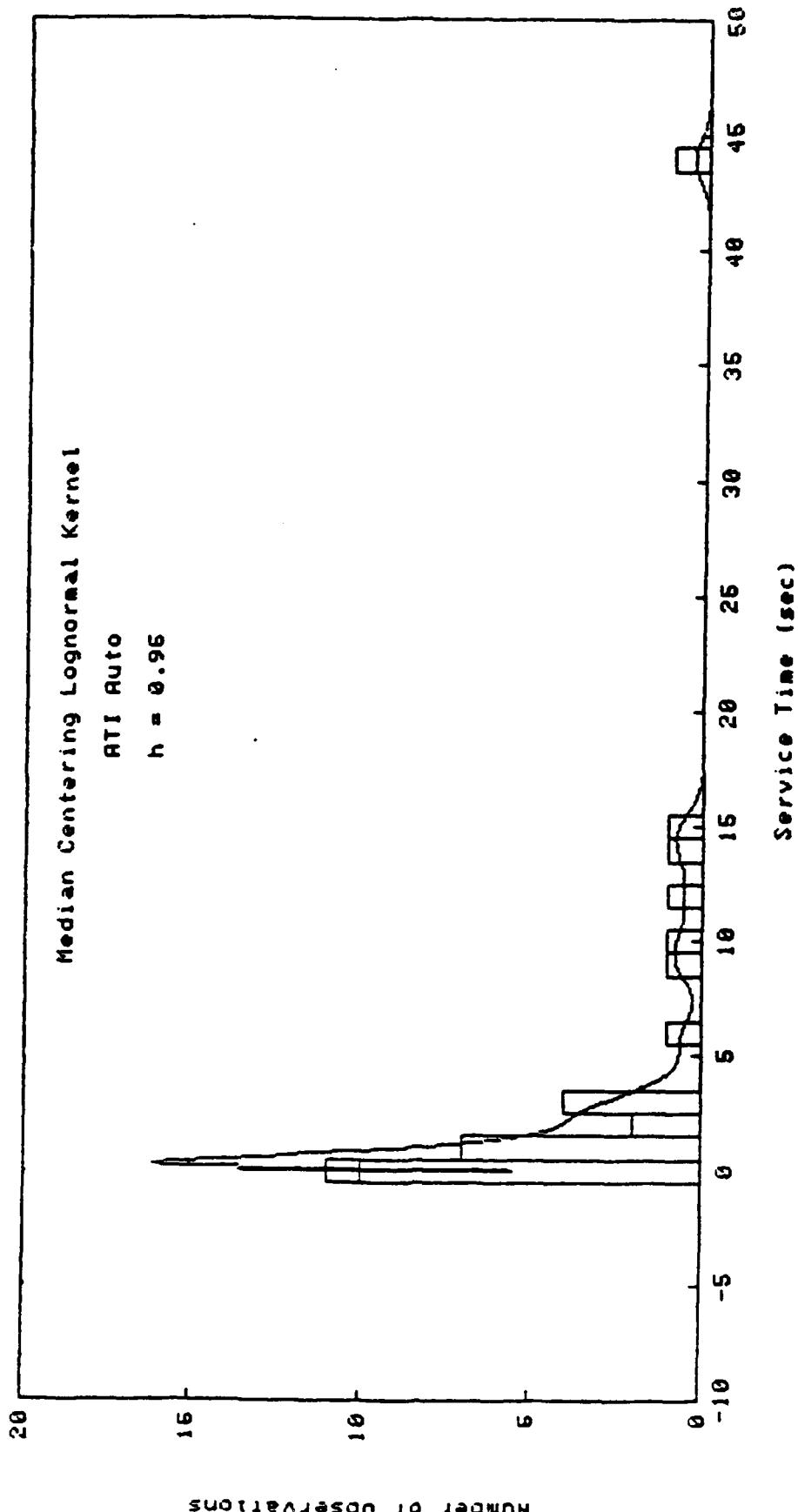
Figure 22. Application of the Median Centering Lognormal Kernel with h=0.95 to ATI Auto Sample.
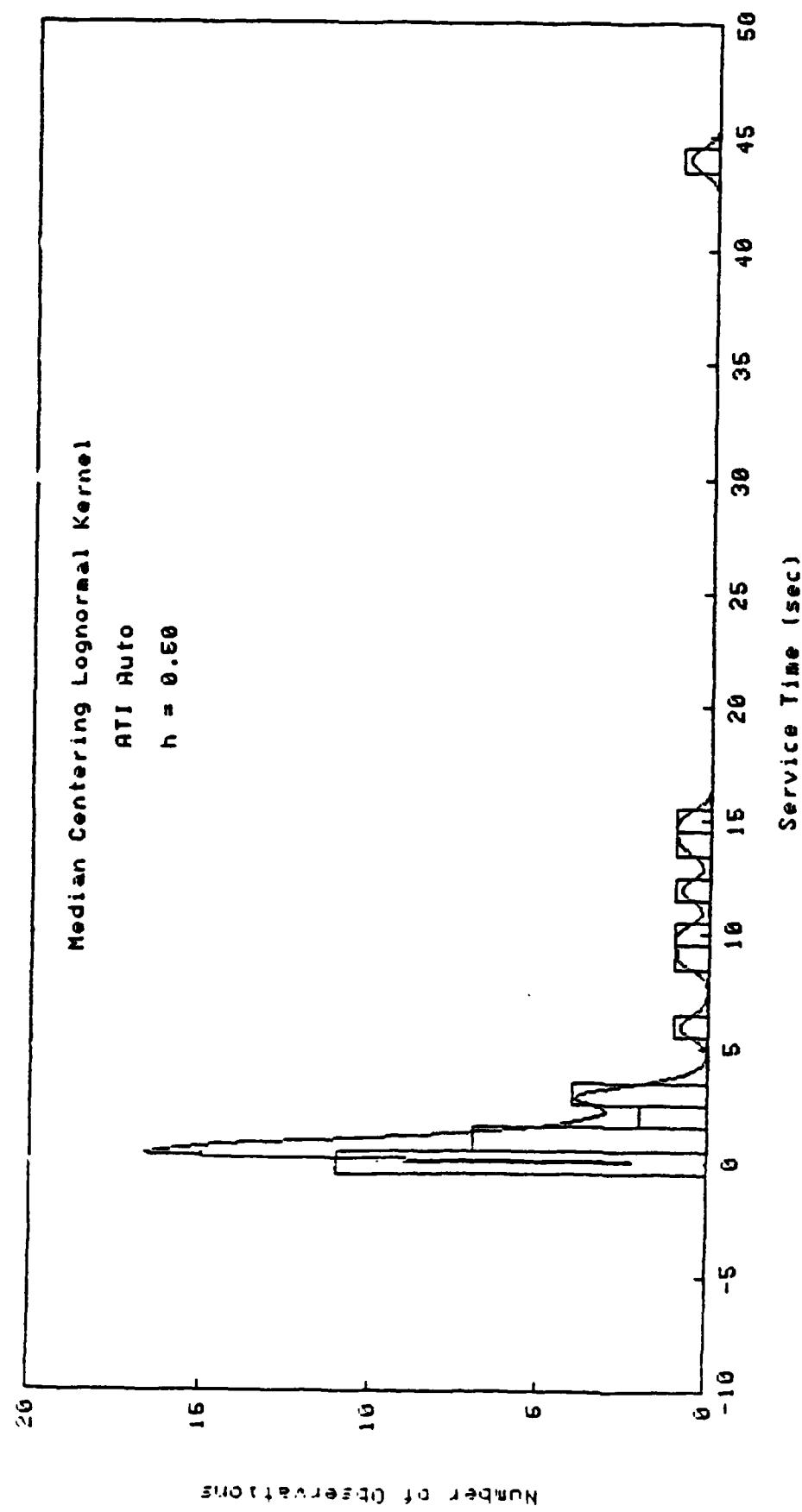
Figure 23. Application of the Median Centering Lognormal Kernel with h=0.50 to ATI Auto Sample.

TABLE 8. Median Centering Lognormal Kernel's Critical h Values
for the FIST HQ Samples

Sample	Number of Modes in interval [0.50]	Critical h Value to two decimal places
ATI Auto	1	9.80
	2	2.26
	3	2.24
	4	0.96
	5	0.92
	6	0.73
	7	0.48
ATI Review	1	9.52
	2	5.03
	3	2.95
	4	2.74
	5	1.43
	6	0.96
	7	0.94

Figures 24 through **26** present kernel density estimates for the ATI Review sample with h equal to 5.00, 2.00 and 0.95, respectively. The density estimate with h equal to 2.00 seems to be a good density estimate and can be compared to the truncated Gaussian density estimate in **Figure 13**. Comparing the oversmoothed median centering Lognormal density estimate in **Figure 24** with the oversmoothed truncated Gaussian kernel density estimate in **Figure 11**, however, shows that the median centering Lognormal kernel produces a concave shaped density estimate, whereas, the truncated Gaussian kernel produces a convex shaped density estimate.

Criticisms of the median centering Lognormal kernel

From the application of the median centering Lognormal kernel to the FIST HQ samples one can see that the median centering Lognormal kernel produces density estimates that are very similar to the ones of the truncated Gaussian kernel. Since many symmetric unimodal kernel functions are nearly optimal. this suggests that. perhaps. many non-symmetric unimodal kernels are also nearly optimal.

VI. CONCLUSIONS

An introduction to kernel density estimation was achieved by examining the application of four different kernels--the Gaussian, truncated Gaussian, mode centering Lognormal and median centering Lognormal kernels--to two FIST FDT&E II samples. This study revealed that each kernel function produced slightly different nonparametric estimates of each sample's underlying probability density function. Monte Carlo simulations with the Gaussian kernel showed that the average efficiency of Tapia and
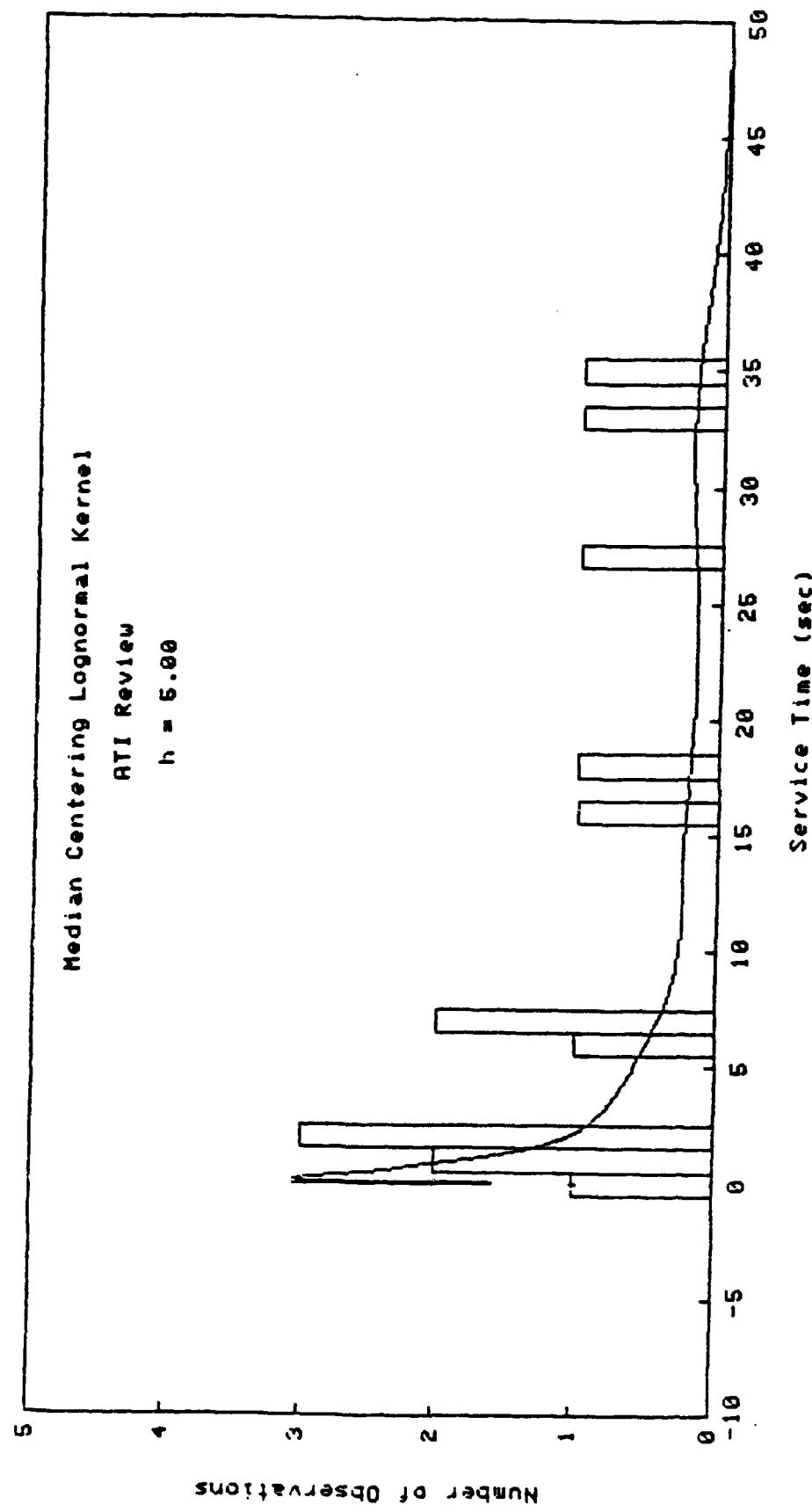
Figure 24. Application of the Median Centering Lognormal Kernel with h=5.00 to ATI Review Sample.
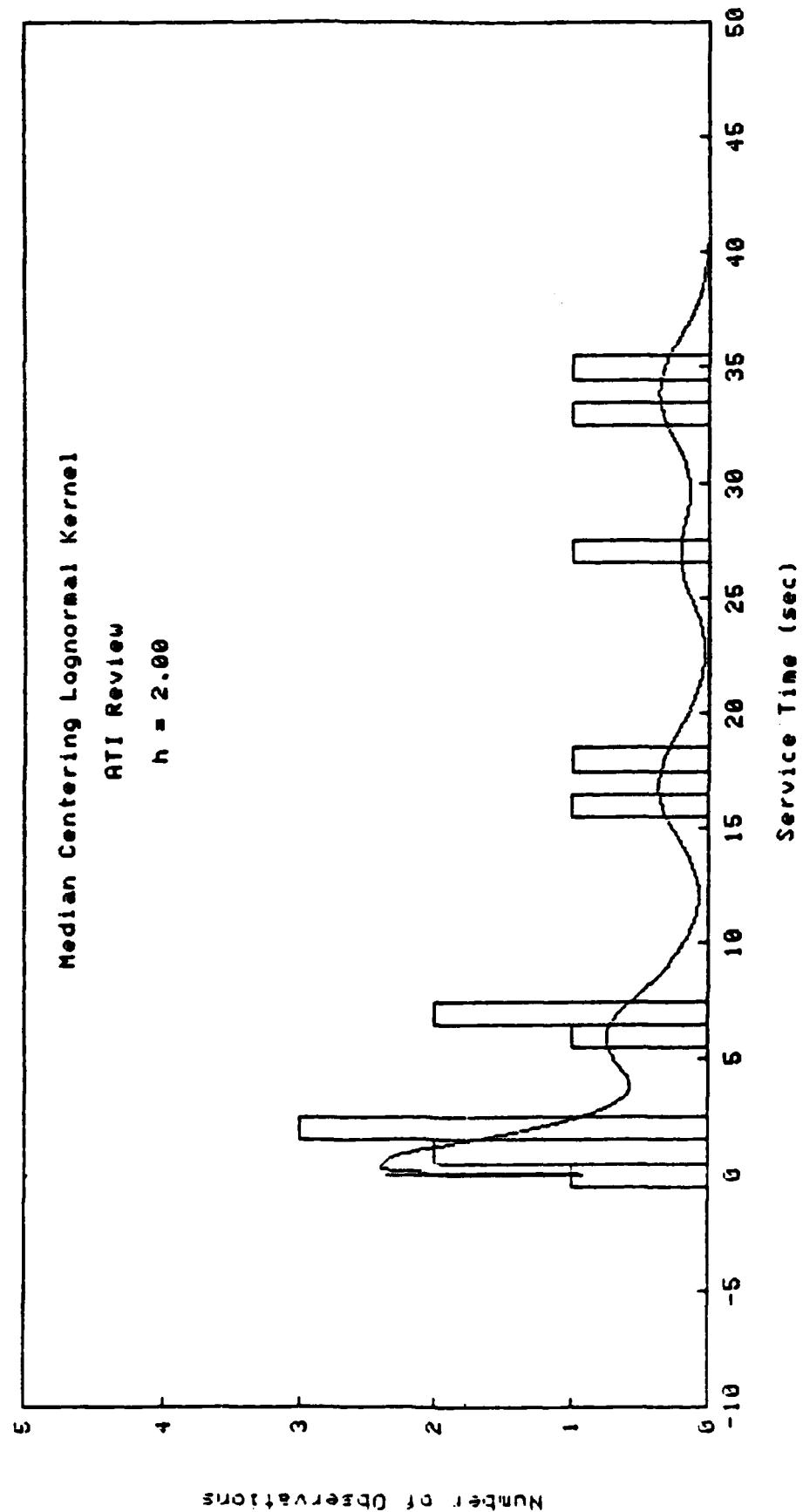
Figure 25. Application of the Median Centering Lognormal Kernel with h=2.00 to ATI Review Sample.

48
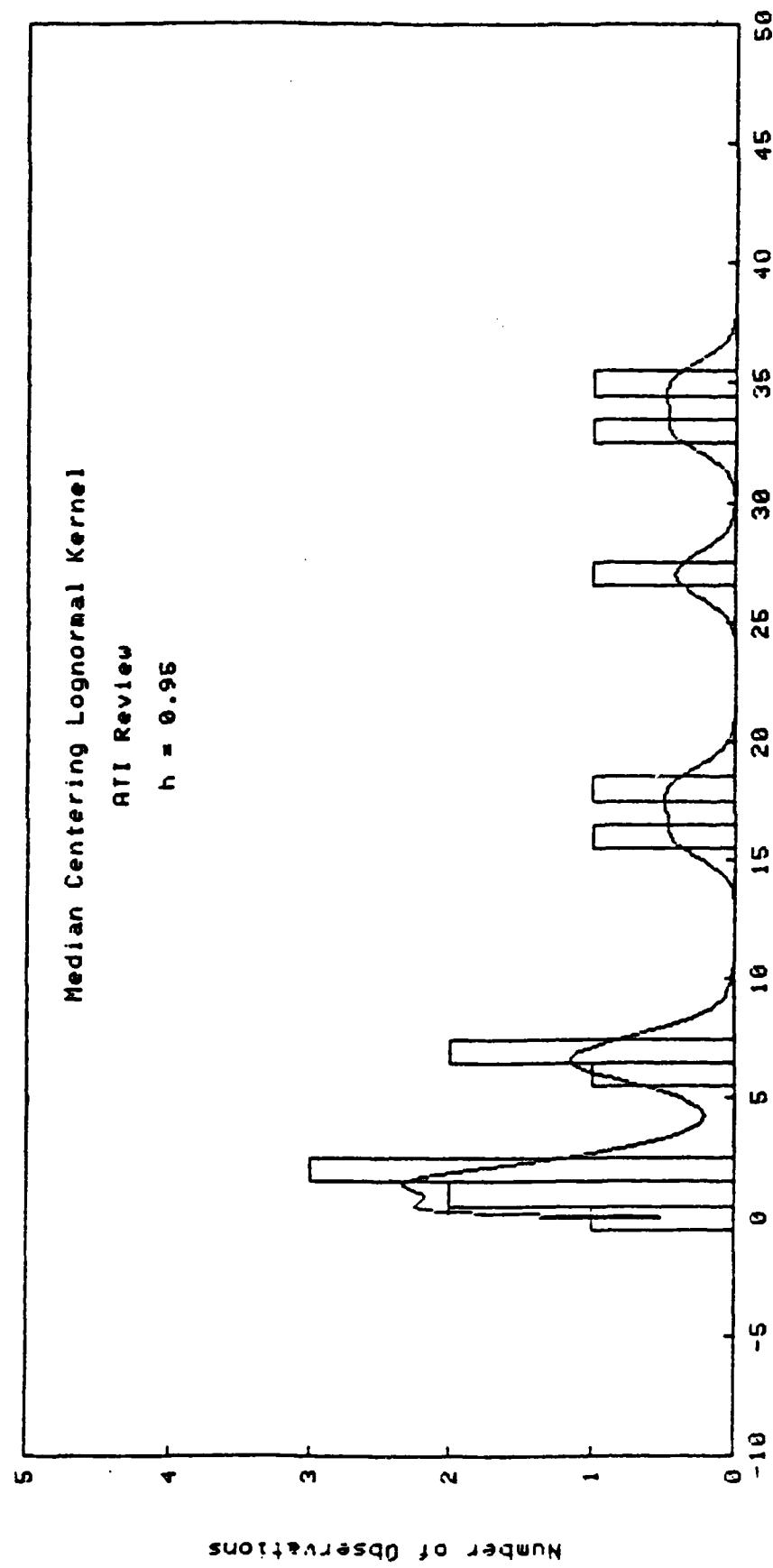
Figure 26. Application of the Median Centering Lognormal Kernel with h=0.95 to ATI Review Sample.

Thompson's formula for this kernel for sample sizes as small as twenty-five was quite good. However, the principle limitation of the Gaussian kernel was the prediction of non-zero densities for negative service time observations. The truncated Gaussian kernel corrected this limitation, and preliminary Monte Carlo simulations suggest that careful selection between these two kernels for a particular sample can increase average efficiency. Applying the median centering Lognormal kernel to the FIST HQ samples resulted in nonparametric density estimates that were very similar to the truncated Gaussian kernel. The mode centering Lognormal kernel produced markedly different density estimates than the three other kernel functions by supplying density estimates that minimized the effect of tail observations and emphasized the observations closest to zero. Tapia & Thompson's recursive formula, which optimizes the smoothing parameter for a sample, enhanced the robustness of Gaussian and truncated Gaussian kernels and provided a quick and efficient starting point for selecting a sample's smoothing parameter value. Additional analysis of the theoretical and experimental causes of the modes suggested by these optimum kernel density estimates should be undertaken to accept the optimum h value or justify a modification of the optimum h value.

From the application of these four different kernel functions to the FIST HQ samples, it was shown that the truncated Gaussian and median centering Lognormal kernels seem to be superior to the Gaussian and mode centering Lognormal kernels for these samples. The principle reasons are due to the following respective inefficiencies of the latter kernels: (1) the prediction of non-zero service time observations, and (2) the sensitivity of the smoothing parameter to small changes. The optimizing kernel density estimates for the truncated Gaussian and median centering Lognormal kernel suggested seven modes for the ATI auto sample and five modes for the ATI review sample in the interval [0,50]. The Gaussian kernel's optimum density estimates also were consistent with these results. By re-examining the histograms accompanying the unnormalized density estimates, one can "see" these predicted modes and observe the relative smoothness of kernel density estimates in comparison to histograms.

In conclusion, this paper showed that kernel density estimation can be useful in producing nonparametric estimates of a sample's underlying probability density function and emphasizes the importance of choosing the appropriate kernel function and smoothing parameter value for a particular sample.

VII. ACKNOWLEDGEMENTS

REFERENCES

1. Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981, p. 10.

2. Kaste, Virginia A., Brodeen, Douglas C. and Winner, Wendy A., "Description of the Digital Data Collected from the FIST FDT&E II," Ballistic Research Laboratory Technical Report No. 2676, 1985.

3. Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, p. 61.

4. Guttman, I. and Wilks, S. S., *Introductory Engineering Statistics*, New York: John Wiley & Sons, 1965 cited by Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981. p. 9.

5. Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, pp. 48,59.

6. Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, p. 60.

7. Parzen, Emanual, "Comment." *Journal of the American Statistical Association*, vol. 75, no. 369, March 1980, p. 59.

8. Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981, pp. 11-12. and Tapia, Richard A. and Thompson, James R., *Nonparametric Probability Density Estimation*, Baltimore, MD: The John Hopkins University Press, 1978, pp. 67-68.

9. Scott, D. W., Tapia, R. A. and Thompson, J. R., "Kernel Density Estimation Revisited," *Journal of Nonlinear Analysis, Theory, Methods and Applications*, vol. 1, 1977 cited by Scott, David W. and Factor, Lynette E., "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, vol. 76, no. 373, March 1981. p. 10.

APPENDIX A. A USER'S GUIDE TO KERNEL PROGRAMS

APPENDIX A. A USER'S GUIDE TO KERNEL PROGRAMS

The purpose of this appendix is to assist the individual who desires to use the programming codes presented in Appendices B through G. These programs are currently available on the U.S. Army Ballistic Research Laboratory's CYBER machine(BRL-CYBER.ARPA). The following instructions describe the files required to run these programs and the specific commands required for the forementioned CYBER. In addition, the instructions for a curve plotting program using DISSPLA Graphics Software Package are also given.

INSTRUCTIONS FOR GAUSSIAN, TRUNCATED GAUSSIAN, MODE CENTERING LOGNORMAL AND MEDIAN CENTERING LOGNORMAL KERNELS

STEP 1.

The user must create and store a MFA permanent file. This file should contain the information listed in **Table A-1**.

TABLE A-1. Data File for Gaussian, Truncated Gaussian, Mode Centering Lognormal and Median Centering Lognormal Kernel

(Note: (PL) denotes program limitation and
* indicates free-format read statements.)

DESCRIPTION	LINE(S)	FORMAT	COLUMNS
1. N—the sample size (PL) $0 < N \leq 1000$ (PL)	1	I10	1-10 -
2. X(I)—an array of N elements containing the sample values	2-1001	REAL	*

STEP 2.

To run these programs in IAF type the following:

BEGIN,PRCKERN,PRCKERN,(1),PRGM=(2)

where

(1) = file name for the file created in STEP 1.

(2) = one of the following program file names

GAUSS {for Gaussian kernel}

TRUNCG {for Truncated Gaussian kernel}

MODLOG {for Mode Centering Lognormal kernel}

MEDLOG {for Median Centering Lognormal kernel}

The user should answer the interactive prompts appropriately and is warned that a crash will occur if prompts are answered inappropriately. The user is also cautioned that if either permanent or local files named PDF or CDF exist prior to entering the BEGIN command, an error 133 message will occur during execution. If this occurs, the user must remove these files and repeat STEP 2.

STEP 3.

After properly completing STEP ⁊, the user has two local and two permanent files called PDF and CDF. File PDF contains the normalized probability density function, and file CDF contains the cumulative distribution function. To obtain a graphical presentation of these files the user should consult INSTRUCTIONS FOR CURVE PLOTTING PROGRAM USING DISSPLA GRAPHICS SOFTWARE PACKAGE.

INSTRUCTIONS FOR THE OPTIMIZING H PROCEDURES FOR THE GAUSSIAN AND TRUNCATED GAUSSIAN KERNELS

STEP 1.

See STEP 1 INSTRUCTIONS FOR GAUSSIAN, TRUNCATED GAUS-
SIAN, MODE CENTERING LOGNORMAL AND MEDIAN CENTERING
LOGNORMAL KERNELS.

STEP 2.

To run these programs in IAF type the following:

BEGIN,PRCOPTH,PRCOPTH,(1),PRGM=(2)

where

(1) = file name for the file created in STEP 1.

(2) = one of the following program file names

GAUSSH {for optimizing h procedure for
        ' the Gaussian kernel} .

TRUNCGH {for optimizing h procedure for
        the truncated Gaussian kernel}

The user should answer the interactive prompts appropriately and is warned
that a crash will occur if prompts are answered inappropriately. The
optimum h value is displayed on the screen during the execution of this pro-
gram.

INSTRUCTIONS FOR CURVE PLOTTING PROGRAM
USING DISSPLA GRAPHICS SOFTWARE PACKAGE

STEP 1.

This program was designed for Tektronics 4014 terminal or a Hewlett Packett 2648 terminal, and subsequently, can only be used in conjunction with these two types of terminals.

STEP 2.

The user must create and store a MFA permanent file. This file should contain the information listed in **Table A-2**. Users may have to alter the call to initialize DISSPLA output to their terminal, and users unfamiliar with DISSPLA are encouraged to consult a DISSPLA manual.

TABLE A-2. Data File for Curve Plotting Program

(NOTE: (PL) denotes program limitation and
* indicates free-format read statements.)

DESCRIPTION	LINE	TYPE/FORMAT	COLUMNS
1. T—the type of terminal being used as follows: 1 for TK4014, 2 for HP2648 and any other integer will will terminate the program.	1	integer	*
2. NPTS—the number of x and y pairs to be plotted	2	integer	*
3. PAGEX, PAGEY—the page size in width and length in inches	3	real	*
4. AXISX, AXISY—the subplot area's width and length in inches	4	real	*
5. LH1—plot heading enclosed in single quotes	5	character	1-60 (PL)
6. H1NUM—total # of characters, letters, spaces, symbols in LH1 (e.g. 7 for 'TITLE') $0 < H1NUM \leq 60$ (PL)	5	integer	*

DESCRIPTION	LINE	TYPE/FORMAT	COLUMNS
7. SIZEH1—LH1's print size	5	real	*
8. NUMHLN—# of headings NUMHLN=3 (PL)	5	integer	*
9. LH2—plot sub-heading enclosed in single quotes	6	character	1-72 (PL)
10. H2NUM—total # of characters, letters, spaces, symbols in LH2 $0 < H2NUM \leq 60$ (PL)	6	integer	*
11. SIZEH2—LH2's print size	6	real	*
12. LH3—plot sub-heading enclosed in single quotes	7	character	1-60 (PL)
13. H3NUM—total # of characters, letters, spaces, symbols in LH3 $0 < H2NUM \leq 60$ (PL)	7	integer	*
14. SIZEH3—LH3's print size	7	real	*
15. LXAXIS—x-axis label enclosed in single quotes	8	character	1-72 (PL)
16. NUMX—total # of characters, letters, spaces, symbols in LXAXIS $0 < NUMX \leq 72$ (PL)	8	integer	*
17. LYAXIS—y-axis label enclosed in single quotes	9	character	1-72 (PL)
18. NUMY—total # of characters, letters, spaces, symbols in LYAXIS $0 < NUMY \leq 72$ (PL)	9	integer	*
19. XORIG—x origin in user's units	10	real	*
20. XSTP—x-axis tick mark unit in user's units	10	real	*
21. XMAX—x maximum value in user's units	10	real	*
22. YORIG—y origin in user's units	10	real	*
23. YSTP—y-axis tick mark unit in user's units	10	real	*
24. YMAX—y maximum value in user's units	10	real	*

STEP 3.

The user should have an MFA permanent normalized probability density function file (i.e., PDF) and/or an MFA permanent cumulative distribution function file (i.e., CDF) created by running one of the programs mentioned in STEP 2's INSTRUCTIONS FOR GAUSSIAN, TRUNCATED GAUSSIAN, MODE CENTERING LOGNORMAL AND MEDIAN CENTERING LOGNORMAL KERNELS.

STEP 4.

To run these programs in IAF using a Hewlett Packard 2648 terminal or a Tektronics 4014 terminal type the following:

BEGIN,PRCDSPL,PRCDSPL,(1),(2)

where

(1) = file name for the file created in STEP 2.

(2) = file name for "PDF" or "CDF" file mentioned in STEP 3.

The DISSPLA plot will be created on the terminal screen. To return to the IAF prompt after the plot is drawn, hit the terminal's return key.

APPENDIX B. GAUSSIAN KERNEL

APPENDIX B. GAUSSIAN KERNEL

```
**********************************************************************
*                          GAUSSIAN KERNEL                           *
*                                                                    *
*  PURPOSE:   THIS PROGRAM COMPUTES THE EMPIRICAL PROBABILITY DENSITY *
*             FUNCTION AND ITS ASSOCIATED CUMULATIVE DISTRIBUTION FUNCTION *
*             FOR THE GAUSSIAN KERNEL.  THIS METHOD IS DESIGNED FOR   *
*             DATA SAMPLES THAT HAVE SAMPLE VALUES RANGING FROM       *
*             NEGATIVE INFINITY TO POSITIVE INFINITY.                 *
*                                                                    *
*  FILES:     THE FOLLOWING DESCRIBES THE INFORMATION REQUIRED BY EACH *
*             FILE.                                                   *
*                                                                    *
*              DATA:        ON THE FIRST LINE MUST BE N, THE SIZE OF  *
*                           THE SAMPLE.  ON THE FOLLOWING LINES OF    *
*                           THE FILE SHOULD BE THE SAMPLE VALUES ONE  *
*                           PER LINE.                                 *
*                                                                    *
*              PDF:         THE NORMALIZED PROBABILITY DENSITY        *
*                           FUNCTION                                  *
*                                                                    *
*              CDF:         THE CUMULATIVE DISTRIBUTION FUNCTION      *
*                                                                    *
**********************************************************************


        PROGRAM GAUSS(DATA,PDF,CDF,INPUT,OUTPUT)

* SPECIFICATION AND DATA STATEMENTS

        INTEGER I, K, N
        REAL ARG, BASE, BOTTOM, F, FHAT, H, MAX, MIN, STEP, X, Z
        DIMENSION X(1000)
        DIMENSION IRAY(6)
        DATA IRAY/6* -0/

* SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
* OF ERROR 115

        IRAY(4)=0
        CALL SYSTEMC(115,IRAY)

* INITIALIZING THE FILES
```

PREVIOUS PAGE
IS BLANK

```
      OPEN(UNIT=1,FILE='DATA',STATUS='OLD')
      REWIND 1
      OPEN(UNIT=2,FILE='PDF',STATUS='NEW')
      REWIND 2
      OPEN(UNIT=3,FILE='CDF',STATUS='NEW')
      REWIND 3
      OPEN(UNIT=5,FILE='INPUT')
      OPEN(UNIT=6,FILE='OUTPUT')

* READING INFORMATION FROM THE DATA FILE

      READ(1,100) N
100   FORMAT(I10)
      MIN = 1.0E10
      MAX = 1.0E-10
      DO 10 I=1,N
         READ(1,*) X(I)
         IF ( X(I) .GT. MAX ) THEN
           MAX = X(I)
         END IF
         IF ( X(I) .LT. MIN ) THEN
           MIN = X(I)
         END IF
10    CONTINUE

* DETERMINING USER REQUIREMENTS

      WRITE(6,'("      GAUSSIAN KERNEL")')
      WRITE(6,101) MIN
101   FORMAT('THE SAMPLE VALUES RANGE FROM ',F15.10)
      WRITE(6,102) MAX
102   FORMAT('TO ',F15.10,'.  INPUT THE FIRST POINT")')
      WRITE(6,'("AT WHICH THE DENSITY ESTIMATE WILL")')
      WRITE(6,'("BE COMPUTED.")')
      READ(5,*) BOTTOM
      WRITE(6,'("INPUT THE LAST POINT AT WHICH")')
      WRITE(6,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
20    READ(5,*) TOP
      IF (TOP .LT. BOTTOM) THEN
        WRITE(6,'("THIS LAST VALUE MUST EXCEED THE FIRST VALUE.")')
        WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
        GOTO 20
      END IF
      WRITE(6,'("INPUT THE SIZE SPACING AT WHICH KERNEL DENSITY")')
      WRITE(6,'("ESTIMATES ARE DESIRED.  FOR EXAMPLE, IF THE")')
      WRITE(6,'("RANGE FOR THE SAMPLE IS 50.0 THEN A SIZE SPACING")')
      WRITE(6,'("EQUAL TO 0.10 WOULD PRODUCE DENSITY ESTIMATES")')
```

```
            WRITE(6,'("AT 0.1, 0.2, 0.3,...,50.0.)")')
    30      READ(5,*) STEP
            IF (STEP .LE. 0) THEN
              WRITE(6,'("SIZE SPACING MUST BE GREATER THAN 0.")')
              WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
              GOTO 30
            END IF
            WRITE(6,'("INPUT A VALUE FOR H, THE SMOOTHING PARAMETER.")')
    40      READ(5,*) H
            IF (H .LE. 0) THEN
              WRITE(6,'("H MUST BE GREATER THAN 0.")')
              WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
              GOTO 40
            END IF


    *  CALCULATING AND STORING F, THE NORMALIZED PROBABILITY DENSITY FUNCTION

            DO 50 Z= BOTTOM, TOP, STEP
              F=0.
              DO 60 I= 1,N
                F = F + EXP(-0.5*((Z-X(I))/H)**2)
    60        CONTINUE
              F = F/SQRT(2.*3.1416)/H/FLOAT(N)
              WRITE(2,103) Z, F
    103       FORMAT(2F15.10)
    50      CONTINUE

    *  NUMERICALLY CALCULATING AND STORING F, THE CUMULATIVE DISTRIBUTION
    *  FUNCTION

            IF (BOTTOM .LT. MIN) THEN
              BASE = BOTTOM
            ELSE
              BASE = MIN
            END IF
    70      FHAT = 0.
            DO 80 I=1,N
              ARG = (BASE - X(I))/H
              FHAT = FHAT + CSND(ARG)
    80      CONTINUE
            IF (FHAT .GT. 0.01) THEN
              BASE = BASE - 1.
              GOTO 70
            END IF
            DO 90 Z= BASE, TOP, STEP
              FHAT=0.
              DO 110 I= 1,N
```

```fortran
          ARG = (Z-X(I))/H
          FHAT = FHAT + CSND(ARG)
110       CONTINUE
      IF (Z .GT. BOTTOM) THEN
        WRITE(3,103) Z, FHAT/FLOAT(N)
      END IF
90    CONTINUE

      END

* THE FOLLOWING FUNCTION APPROXIMATES THE INTEGRAL FROM NEGATIVE
* INFINITY TO Y OF  (1/SQRT(2*PI)) EXP(-1*(T**2)/2) DT (I.E., THE
* CUMULATIVE STANDARD NORMAL DISTRIBUTION).

      FUNCTION CSND(X)
        Y = ABS(X)
        P=(((((5.383E-06*Y+4.88906E-05)*Y+3.80036E-05)*Y+
     &      0.0032776263)*Y+0.0211410061)*Y+0.0498673469)*Y
        P=0.5*(P+1.)**(-16)
        IF (X .GT. 0.) THEN
          CSND=1.-P
        ELSE
          CSND=P
        END IF
        RETURN
      END
```

68

APPENDIX C. OPTIMIZING H PROCEDURE FOR THE GAUSSIAN KERNEL

APPENDIX C. OPTIMIZING H PROCEDURE FOR THE GAUSSIAN KERNEL

```
********************************************************************
*                OPTIMIZING H PROCEDURE FOR GAUSSIAN KERNEL        *
*                                                                  *
* PURPOSE:  THE PURPOSE OF THIS PROGRAM IS TO OBTAIN A SATISFACTORY VALUE *
*           FOR THE GAUSSIAN KERNEL'S H VALUE.                     *
*                                                                  *
* FILES:    BELOW IS A BRIEF SYNOPSIS OF THE EXPECTED CONTENTS FOR THE FILE *
*           DATA.                                                  *
*                                                                  *
*           DATA:  ON THE FIRST LINE OF THIS FILE SHOULD BE THE NUMBER *
*                  OF SAMPLE VALUES CONTAINED IN THIS FILE.  ON THE *
*                  FOLLOWING LINES SHOULD BE THE SAMPLE VALUES ONE PER *
*                  LINE.                                           *
********************************************************************

        PROGRAM GAUSSH(DATA, INPUT, OUTPUT)

* SPECIFICATION AND DATA STATEMENTS

        INTEGER I, N, TOTAL
        REAL ALPHA, BETA, C, DIFF, H, NEWH, X
        DIMENSION X(1000)
        DIMENSION IRAY(6)
        DATA IRAY/6* -0/

* SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
* OF ERROR 115

        IRAY(4)=0
        CALL SYSTEMC(115, IRAY)

* OPENING THE FILES

        OPEN(UNIT=1, FILE='DATA', STATUS='OLD')
        REWIND 1
        OPEN(UNIT=5, FILE='INPUT')
        OPEN(UNIT=6, FILE='OUTPUT')

* SETTING THE CONSTANTS.

        ALPHA = ((1./(2.*SQRT(3.1416)))**(0.2))
        TOTAL = 1000

* READING FROM THE FILE DATA
```

PREVIOUS PAGE
IS BLANK

```fortran
          CALL    ALUES(X,N,H,DIFF)

* EMPLOYING THE OPTIMIZING PROCEDURE OUTLINED BY RICHARD A. TAPIA AND
* JAMES R. THOMPSON IN THEIR BOOK NONPARAMETRIC PROBABILITY DENSITY
* ESTIMATION, BALTIMORE, MD: THE JOHN HOPKINS UNIVERSITY PRESS, 1978,
* P. 59.

          WRITE(6,'("GUESS FOR H:")')
          WRITE(6,100) H
  100     FORMAT(F15.5)
          WRITE(6,'("--------------")')
          C = N**(-0.2)
          DO 10 I= 1, TOTAL
            CALL BETAF(X, BETA, N, H)
            NEWH = C * ALPHA * BETA
            IF (NEWH .LT. 0.01) THEN
              WRITE(6,'("H:")')
              WRITE(6,'("DIRAC SPIKE")')
              GOTO 20
            END IF
            IF ( ABS(NEWH-H) .LT. DIFF ) THEN
              WRITE(6,101) DIFF
  101         FORMAT(6,'AN ACCEPTABLE H WITHIN ',F15.10,' IS')
              WRITE(6,100) NEWH
              GOTO 20
            END IF
            WRITE(6,'("H:")')
            WRITE(6,100) NEWH
            H = NEWH
  10      CONTINUE

  20      END


* THIS SUBROUTINE READS THE VALUES FROM THE FILE DATA AND INPUT.

          SUBROUTINE VALUES(X,N,H,DIFF)

          REAL DIFF, H, MIN, MAX
          DIMENSION X(1000)

          READ(1,*) N
          MAX = 1.0E-10
          MIN = 1.0E10
          DO 30 I= 1, N
            READ(1,*) X(I)
            IF (X(I) .GT. MAX) THEN
```

```
                    MAX = X(I)
                 END IF
                 IF (X(I) .LT. MIN) THEN
                    MIN = X(I)
                 END IF
30            CONTINUE

              WRITE(6,'("  OPTIMIZING H VALUE FOR THE GAUSSIAN KERNEL")')
              WRITE(6,'(" ")')
              WRITE(6,'("CHOOSING THE INITIAL H VALUE TO BE THE SAMPLE")')
              WRITE(6,'("RANGE GUARANTEES THE CONVERGENCE OF THIS ITERA-")')
              WRITE(6,'("TIVE SEQUENCE TO THE LARGEST NON-NEGATIVE")')
              WRITE(6,102) MAX - MIN
102           FORMAT('SOLUTION.  THE RANGE OF THIS SAMPLE IS ',F15.10,'.')
              WRITE(6,'("INPUT A VALUE FOR THE INITIAL H VALUE.")')
              READ(5,*) H
              WRITE(6,'("INPUT THE MINIMUM ACCEPTABLE DIFFERENCE BETWEEN")')
              WRITE(6,'("THE SUCCESSIVE H VALUES.  A RECOMMENDED DIFFER-")')
              WRITE(6,'("ENCE IS 1.0E-5")')
              READ(5,*) DIFF
           END




* THIS SUBROUTINE APPROXIMATES TAPIA AND THOMPSON'S BETA VALUE
* BY USING THE FORMULA OUTLINED BY SCOTT AND FACTOR IN THE JOURNAL
* OF THE AMERICAN STATISTICAL ASSOCIATION, VOL. 76, NO. 373, MARCH
* 1981, P. 10.

           SUBROUTINE BETAF(X,BETA,N,H)

              REAL  HF, HS
              DIMENSION X(1000)

              A = 1./12.
              HS = H**2
              HF = HS**2
              BETA = 0.
              DO 40 J=1,N
                DO 50 K = 1,N
                   Y = X(J) - X(K)
                   Z = EXP(-(Y**2)/(4.*HS))
                   BETA = BETA + (HF - ((Y**2)*HS) + (A*(Y**4)) )*Z
50              CONTINUE
40            CONTINUE
              BETA = (BETA*(3./(8.*SQRT(3.1416)*N*N*(H**9))))**(-1./5.)
           END
```

APPENDIX D. TRUNCATED GAUSSIAN KERNEL

APPENDIX D. TRUNCATED GAUSSIAN KERNEL

```
********************************************************************
*                  TRUNCATED GAUSSIAN KERNEL                       *
*                                                                  *
*                                                                  *
* PURPOSE:   THIS PROGRAM COMPUTES THE NORMALIZED PROBABILITY      *
*            DENSITY FUNCTION AND ITS ASSOCIATED CUMULATIVE DISTRIBUTION *
*            FUNCTION BY USING A TRUNCATED GAUSSIAN KERNEL FOR THOSE *
*            DATA SAMPLE VALUES LYING BELOW 3.*H AND THE TYPICAL GAUSSIAN *
*            KERNEL VALUES FOR THE DATA SAMPLE VALUES ABOVE 3.*H.   THIS  *
*            DESIGN IS CHOSEN FOR DATA SAMPLES THAT HAVE NON-NEGATIVE *
*            SAMPLE VALUES AND CAN THEORETICALLY ONLY HAVE NON-NEGATIVE *
*            VALUES.                                               *
*                                                                  *
* FILES:     THE FOLLOWING DESCRIBES THE INFORMATION IN EACH FILE.  *
*                                                                  *
*                  DATA:       ON THE FIRST LINE MUST BE THE SIZE OF THE *
*                              SAMPLE, N.   ON THE FOLLOWING LINES OF THE *
*                              FILE SHOULD BE THE SAMPLE VALUES ONE PER *
*                              LINE.                               *
*                                                                  *
*                  PDF:        THE NORMALIZED PROBABILITY DENSITY   *
*                              FUNCTION                            *
*                                                                  *
*                  CDF:        THE CUMULATIVE DISTRIBUTION FUNCTION  *
********************************************************************

        PROGRAM TRUNCG(DATA,PDF,CDF,INPUT,OUTPUT)

* SPECIFICATION AND DATA STATEMENTS

        INTEGER I, K, N
        REAL ARG, BASE, BOTTOM, F, H, MAX, MIN, STEP, TOP, X ,Z
        DIMENSION X(1000,2)
        DIMENSION IRAY(6)
        DATA IRAY/6* -0/

* SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
* OF ERROR 115

        IRAY(4)=0
        CALL SYSTEMC(115,IRAY)

* INITIALIZING THE FILES

        OPEN(UNIT=1,FILE='DATA',STATUS='OLD')
        REWIND 1
```

```fortran
      OPEN(UNIT=2,FILE='PDF',STATUS='NEW')
      REWIND 2
      OPEN(UNIT=3,FILE='CDF',STATUS='NEW')
      REWIND 3
      OPEN(UNIT=5,FILE='INPUT')
      OPEN(UNIT=6,FILE='OUTPUT')

* READING INFORMATION FROM THE INPUT FILES AND CALCULATING X(I,2)

      WRITE(6,'("      TRUNCATED GAUSSIAN KERNEL")')
      WRITE(6,'(" ")')
      WRITE(6,'("INPUT H, THE SMOOTHING PARAMETER.")')
10    READ(5,*) H
      IF (H .LE. 0) THEN
        WRITE(6,'("H MUST BE GREATER THAN 0.")')
        WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
        GOTO 10
      END IF

      READ(1,101) N
101   FORMAT(I10)
      MAX = 1.0E-10
      DO 20 I=1,N
        READ(1,*) X(I,1)
        IF ( X(I,1) .GT. 3.*H ) THEN
          X(I,2) = 1.
          IF ( X(I,1) .GT. MAX) THEN
            MAX = X(I,1)
          END IF
        ELSE
          ARG = X(I,1)/H
          X(I,2) = CSND(ARG)
        END IF
20    CONTINUE

* DETERMINING USER REQUIREMENTS

      WRITE(6,102) MIN
102   FORMAT('THE SAMPLE VALUES RANGE FROM ',F15.10)
      WRITE(6,103) MAX
103   FORMAT('TO ',F15.10,'.  INPUT THE FIRST')
      WRITE(6,'("NON-NEGATIVE POINT AT WHICH THE DENSITY")')
      WRITE(6,'("ESTIMATE WILL BE COMPUTED.")')
30    READ(5,*) BOTTOM
      IF (BOTTOM .LT. 0) THEN
        WRITE(6,'("THIS VALUE MUST BE NON-NEGATIVE.")')
        WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
```

```fortran
            GOTO 30
          END IF
          WRITE(6,'("INPUT THE LAST NON-NEGATIVE POINT AT WHICH")')
          WRITE(6,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
   40     READ(5,*) TOP
          IF (TOP .LT. BOTTOM) THEN
            WRITE(6,'("THIS LAST VALUE MUST EXCEED THE FIRST VALUE.")')
            WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
            GOTO 40
          END IF
          WRITE(6,'("INPUT THE SIZE SPACING AT WHICH KERNEL DENSITY")')
          WRITE(6,'("ESTIMATES ARE DESIRED.  FOR EXAMPLE, IF THE")')
          WRITE(6,'("RANGE FOR THE SAMPLE IS 50.0 THEN A SIZE SPACING")')
          WRITE(6,'("EQUAL TO 0.10 WOULD PRODUCE DENSITY ESTIMATES")')
          WRITE(6,'("AT 0.1, 0.2, 0.3,...,50.0.)")')
   50     READ(5,*) STEP
          IF (STEP .LE. 0) THEN
            WRITE(6,'("SIZE SPACING MUST BE GREATER THAN 0.")')
            WRITE(6,'("PLEASE INPUT A NEW VALUE.")')
            GOTO 50
          END IF


* CALCULATING AND STORING F, THE NORMALIZED PROBABILITY DENSITY FUNCTION

          DO 60 Z= BOTTOM, TOP, STEP
            F=0.
            DO 70 I= 1,N
              F= F + EXP(-0.5*((Z-X(I,1))/H)**2)/X(I,2)
   70       CONTINUE
            F = F/SQRT(2.*3.1416)/H/FLOAT(N)
            WRITE(2,104) Z, F
  104       FORMAT(2F15.10)
   60     CONTINUE


* NUMERICALLY CALCULATING AND STORING F, THE CUMULATIVE DISTRIBUTION FUNCTION

          IF (BOTTOM .LT. MIN) THEN
            BASE = BOTTOM
          ELSE
            BASE = MIN
          END IF
   80     FHAT = 0.
          DO 90 I=1,N
            ARG = (BASE - X(I,1))/H
            FHAT = FHAT + CSND(ARG)
   90     CONTINUE
          IF (FHAT .GT. 0.01) THEN
```

79

```fortran
         IF ( (BASE-1.) .LT. 0.) THEN
           BASE = 0.
           GOTO 100
         ELSE
           BASE = BASE - 1.
           GOTO 80
         END IF
       END IF
100      DO 110 Z= BASE, TOP, STEP
         F=0.
         DO 120 I= 1,N
           ARG = (Z-X(I,1))/H
           IF (X(I,2) .NE. 1) THEN
             F = F + ((CSND(ARG)-1)/X(I,2)) + 1
           ELSE
             F = F + CSND(ARG)
           END IF
120        CONTINUE
         IF (Z .GT. BOTTOM) THEN
           WRITE(3,104) Z, F/FLOAT(N)
         END IF
110      CONTINUE
       END


*  THE FOLLOWING FUNCTION APPROXIMATES THE INTEGRAL FROM NEGATIVE
*  INFINITY TO Y OF (1/SQRT(2*PI)) EXP(-1*(T**2)/2) DT (I.E., THE
*  CUMULATIVE STANDARD NORMAL DISTRIBUTION).

       FUNCTION CSND(X)
         Y = ABS(X)
         P=(((((5.383E-06*Y+4.88906E-05)*Y+3.80036E-05)*Y+
     &       0.0032776263)*Y+0.0211410061)*Y+0.0498673469)*Y
         P=0.5*(P+1.)**(-16)
         IF (X .GT. 0.) THEN
           CSND=1.-P
         ELSE
           CSND=P
         END IF
         RETURN
       END
```

APPENDIX E. OPTIMIZING H PROCEDURE FOR THE TRUNCATED GAUSSIAN KERNEL

APPENDIX E. OPTIMIZING H PROCEDURE FOR THE TRUNCATED GAUSSIAN KERNEL

```
********************************************************************
*                OPTIMIZING H PROCEDURE FOR TRUNCATED GAUSSIAN KERNEL    *
*                                                                  *
* PURPOSE:   THE PURPOSE OF THIS PROGRAM IS TO OBTAIN A SATISFACTORY VALUE *
*            FOR THE TRUNCATED GAUSSIAN KERNEL'S H.   THE RECURSIVE METHOD *
*            EMPLOYED IS OUTLINED BY RICHARD A. TAYLOR AND JAMES R. THOMPSON *
*            ON PP. 57-67 OF THEIR BOOK NONPARAMETRIC PROBABILITY DENSITY *
*            ESTIMATION (BALTIMORE, MD: THE JOHN HOPKINS UNIVERSITY PRESS, *
*            1978).                                                *
*                                                                  *
* FILES:     BELOW IS A BRIEF SYNOPSIS OF THE EXPECTED CONTENTS FOR THE *
*            FILE DATA.                                            *
*                                                                  *
*                    DATA:  ON THE FIRST LINE OF THIS FILE SHOULD BE THE NUMBER *
*                           OF SAMPLE VALUES CONTAINED IN THIS FILE.   ON THE *
*                           FOLLOWING LINES SHOULD BE THE SAMPLE VALUES ONE PER *
*                           LINE.                                  *
********************************************************************

          PROGRAM TRUNCGH(DATA, INPUT, OUTPUT)

* SPECIFICATION AND DATA STATEMENTS

          INTEGER K, N, TOTAL
          REAL AB, BOTTOM, C, DIFF, F, H, NEWH, STEP, TOP, Z
          DIMENSION X(1000), Z(1000), F(1000)
          DIMENSION IRAY(6)
          DATA IRAY/6* -0/

* SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
* OF ERROR 115

          IRAY(4)=0
          CALL SYSTEMC(115, IRAY)

* OPENING THE FILES

          OPEN(UNIT=1, FILE='DATA', STATUS='OLD')
          REWIND 1
          OPEN(UNIT=5, FILE='INPUT')
          OPEN(UNIT=6, FILE='OUTPUT')

* SETTING THE CONSTANTS.

          TOTAL = 50
```

PREVIOUS PAGE
IS BLANK

```
          K = 500


• READING FROM THE FILES

          CALL VALUES(X,N,H,K,BOTTOM,TOP,STEP,DIFF)


• EMPLOYING THE OPTIMIZING H PROCEDURE OUTLINED BY RICHARD A. TAPIA
• AND JAMES R. THOMPSON IN THEIR BOOK NONPARAMETRIC PROBABILITY DENSITY
• ESTIMATION, BALTIMORE, MD: THE JOHN HOPKINS UNIVERSITY PRESS. 1978,
• P. 59.

          WRITE(6,'("GUESS FOR H:")')
          WRITE(6,101) H
  101     FORMAT(F15.5)
          C = N**(-0.2)
          DO 10 I= 1, TOTAL
            CALL ABF(X, Z, F, AB, N, H, K, BOTTOM, TOP, STEP)
            NEWH = C * AB
            IF ( ABS(NEWH-H) .LT. DIFF ) THEN
              WRITE(6,100) DIFF
  100         FORMAT('AN ACCEPTABLE H WITHIN ',F15.10,' IS')
              WRITE(6,101) NEWH
              GOTO 20
            END IF
            IF (NEWH .LT. DIFF) THEN
              WRITE(6,'("DIRAC SPIKE")')
              GOTO 20
            END IF
            WRITE(6,'("H:")')
            WRITE(6,101) NEWH
            H = NEWH
  10      CONTINUE

  20      END



• THIS SUBROUTINE READS THE VALUES FROM THE FILE DATA AND INPUT.

          SUBROUTINE VALUES(X,N,H,K,BOTTOM,TOP,STEP,DIFF)

          REAL MIN, MAX
          DIMENSION X(1000)

          READ(1,102) N
  102     FORMAT(I10)
          MIN = 1.0E10
          MAX = 1.0E-10
```

34

```fortran
         DO 30 I= 1, N
           READ(1,103) X(I)
103        FORMAT(F10.0)
           IF ( X(I) .GT. MAX ) THEN
             MAX = X(I)
           END IF
           IF ( X(I) .LT. MIN ) THEN
             MIN = X(I)
           END IF
30         CONTINUE
         IF ( MIN .LT. 0.) THEN
           MIN = 0.
         END IF

         WRITE(6,'("OPTIMIZING H FOR THE TRUNCATED GAUSSIAN KERNEL")')
         WRITE(6,'(" ")')
         WRITE(6,'("CHOOSING THE INITIAL H VALUE TO BE THE SAMPLE")')
         WRITE(6,'("RANGE GUARANTEES THE CONVERGENCE OF THIS ITERA-")')
         WRITE(6,'("TIVE SEQUENCE TO THE LARGEST NON-NEGATIVE")')
         WRITE(6,'("SOLUTION. THE RANGE OF THIS SAMPLE IS")')
         WRITE(6,*) MAX - MIN
         WRITE(6,'("INPUT A GUESS FOR H, THE SMOOTHING PARAMETER.")')
         READ(5,*) H
         WRITE(6,'("INPUT THE MINIMUM ACCEPTABLE DIFFERENCE BETWEEN")')
         WRITE(6,'("THE SUCCESSIVE H VALUES.  A RECOMMENDED DIFFER-")')
         WRITE(6,'("ENCE IS 1.0E-5")')
         READ(5,*) DIFF
         WRITE(6,'("THE ITERATIVE H VALUES ARE")')
         WRITE(6,'(" ")')
         BOTTOM = MIN
         TOP = MAX + 3.*H
         STEP = ( TOP - BOTTOM )/FLOAT(K)

         END




* THIS SUBROUTINE APPROXIMATES TAPIA AND THOMPSON'S ALPHA*BETA VALUE
* (I.E., AB).

         SUBROUTINE ABF(X,Z,F,AB,N,H,K,BOTTOM,TOP,STEP)

         DIMENSION X(1000), Z(1000), F(1000)

         ALPHA1 = (0.7764)**(-2.5)
         ALPHA2 = (0.8918)**(-2.5)
         A1 = ALPHA1/(SQRT(2.*3.1416)*(H**3)*N)
```

85

```
            A2 = ALPHA2*SQRT(0.5)/(SQRT(3.1416)*N*(H**2))
            I = 1
            DO 40 R= BOTTOM, TOP, STEP
              F(I) = 0.
              Z(I) = R
              DO 50 J = 1,N
                Y = (Z(I) - X(J))/H
                IF (X(J) .GE. (3.*H)) THEN
                  F(I) = F(I) + A1 * EXP(-0.5*Y*Y) * (Y*Y - 1.)
                ELSE
                  F(I) = F(I) + A2 * EXP(-0.5*Y*Y)* (Y*Y - 1.) * 2.
                END IF
50            CONTINUE
              F(I) = F(I)*F(I)
              I = I + 1
40          CONTINUE

*     CALCULATING THE BETA VALUE OF AB BY SIMPSON'S 1/3 RULE OVER 2
*     SUBINTERVALS

            AB = 0.
            S = STEP/3.
            .DO 60 I = 1, K-2, 2
              AB = AB + S * (F(I) + (4.*F(I+1)) + F(I+2))
60          CONTINUE
            AB = AB**(-0.2)

        END
```

APPENDIX F. MODE CENTERING LOGNORMAL KERNEL

APPENDIX F. MODE CENTERING LOGNORMAL KERNEL

```
••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••
•                   MODE CENTERING LOGNORMAL KERNEL                •
•                                                                 •
• PURPOSE:  THIS PROGRAM WILL COMPUTE THE PROBABILITY DENSITY      •
•           FUNCTION AND ITS ASSOCIATED CUMULATIVE DISTRIBUTION FUNCTION •
•           BY USING A MODE CENTERING LOGNORMAL KERNEL.  DUE TO THE MATHE-•
•           MATICAL CONSTRAINTS OF THE LOGNORMAL KERNEL ALL DATA SAMPLE  •
•           VALUES MUST BE STRICTLY POSITIVE.                      •
•                                                                 •
• FILES:    BELOW IS A BRIEF DESCRIPTION OF EACH FILE'S EXPECTED CONTENTS.•
•                                                                 •
•              DATA:      ON THE FIRST LINE OF THIS FILE SHOULD BE THE •
•                         SAMPLE SIZE, N.  ON THE FOLLOWING LINES SHOULD •
•                         BE THE SAMPLE VALUES ONE PER LINE.       •
•                                                                 •
•              PDF:       THE NORMALIZED PROBABILITY DENSITY FUNCTION •
•                                                                 •
•              CDF:       THE CUMULATIVE DISTRIBUTION FUNCTION     •
••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••


         PROGRAM MODELOG(DATA,PDF,CDF)


• SPECIFICATION AND DATA STATEMENTS

         INTEGER I, J, K, N
         REAL BOTTOM, F, FHAT, H, MAX, MIN, R, S, STEP, TOP, X, Z
         DIMENSION F(1000), X(1000), Z(1000)
         DIMENSION IRAY(6)
         DATA IRAY/6* -0/


• SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
• OF ERROR 115

         IRAY(4)=0
         CALL SYSTEMC(115,IRAY)


• OPENING THE FILES

         OPEN(UNIT=1,FILE='DATA',STATUS='OLD')
         REWIND 1
         OPEN(UNIT=2,FILE='PDF',STATUS='NEW')
         REWIND 2
         OPEN(UNIT=3,FILE='CDF',STATUS='NEW')
         REWIND 3
```

• READING DATA FROM INPUT FILES

```
        READ(1,100) N
100     FORMAT(I15)
        MAX = 0.
        MIN = 0.
        DO 10 I=1,N
          READ(1,*) X(I)
          IF ( X(I) .GT. MAX ) THEN
            MAX = X(I)
          ELSE
            IF ( X(I) .LT. MIN) THEN
              MIN = X(I)
            END IF
          END IF
10      CONTINUE
```

• DETERMINING USER REQUIREMENTS

```
        WRITE(*,'("THE SAMPLE VALUES RANGE FROM")')
        WRITE(*,*) MIN
        WRITE(*,'("TO")')
        WRITE(*,*) MAX
        WRITE(*,'("INPUT THE FIRST NON-NEGATIVE POINT AT WHICH")')
        WRITE(*,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
15      READ(*,*) BOTTOM
        IF (BOTTOM .LE. 0) THEN
          WRITE(*,'("THIS VALUE MUST BE GREATER THAN ZERO.")')
          WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
          GOTO 15
        END IF
        WRITE(*,'("INPUT THE LAST NON-NEGATIVE POINT AT WHICH")')
        WRITE(*,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
25      READ(*,*) TOP
        IF (TOP .LT. BOTTOM) THEN
          WRITE(*,'("THIS LAST VALUE MUST EXCEED THE FIRST VALUE.")')
          WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
          GOTO 25
        END IF
        WRITE(*,'("INPUT THE SIZE SPACING AT WHICH KERNEL DENSITY")')
        WRITE(*,'("ESTIMATES ARE DESIRED.  FOR EXAMPLE, IF THE")')
        WRITE(*,'("RANGE FOR THE SAMPLE IS 50.0 THEN A SIZE SPACING")')
        WRITE(*,'("EQUAL TO 0.10 WOULD PRODUCE DENSITY ESTIMATES")')
        WRITE(*,'("AT 0.1, 0.2, 0.3,...,50.0.)")')
35      READ(*,*) STEP
        IF (STEP .LE. 0) THEN
          WRITE(*,'("SIZE SPACING MUST BE GREATER THAN 0.")')
```

```
                    WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
                    GOTO 35
                END IF
                WRITE(*,'("INPUT A VALUE FOR H, THE SMOOTHING PARAMETER.")')
    45          READ(*,*) H
                IF (H .LE. 0) THEN
                    WRITE(*,'("H MUST BE GREATER THAN 0.")')
                    WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
                    GOTO 45
                END IF

* CALCULATING AND STORING F, THE NORMALIZED PROBABILITY DENSITY FUNCTION

                I = 1
                DO 20 R= STEP, TOP, STEP
                    F(I) =0.0
                    DO 30 J= 1,N
                        A = (LOG(R)-H*H-LOG(X(J)))**2
                        F(I)= F(I) + EXP(-0.5*A/H**2)
    30          CONTINUE
                F(I) = F(I)/(SQRT(2.*3.1416) * H * R * FLOAT(N) )
                Z(I) = R
                WRITE(2,102) Z(I), F(I)
    102         FORMAT(2F15.10)
                I = I + 1
    20      CONTINUE

* NUMERICALLY CALCULATING AND STORING FHAT, THE CUMULATIVE DISTRIBUTION
* FUNCTION, BY USING SIMPSON'S 1/3 RULE OVER 2 SUBINTERVALS

                K = INT((TOP - BOTTOM)/STEP)
                S = ABS(Z(K-1)/K)/3.
                FHAT = 0.0
                DO 40 I= 1, K-2, 2
                    FHAT = FHAT + (S * (F(I) + (4*F(I+1)) + F(I+2)))
                    WRITE(3,102) Z(I), FHAT
    40      CONTINUE
                END
```

APPENDIX G. MEDIAN CENTERING LOGNORMAL KERNEL

APPENDIX G. MEDIAN CENTERING LOGNORMAL KERNEL

```
*******************************************************************
*                MEDIAN CENTERING LOGNORMAL KERNEL                *
*                                                                 *
* PURPOSE:  THIS PROGRAM WILL COMPUTE THE PROBABILITY DENSITY      *
*           FUNCTION AND ITS ASSOCIATED CUMULATIVE DISTRIBUTION FUNCTION *
*           BY USING A MEDIAN CENTERING LOGNORMAL KERNEL.  DUE TO THE    *
*           MATHEMATICAL CONSTRAINTS OF THE LOGNORMAL KERNEL ALL DATA    *
*           SAMPLE VALUES MUST BE STRICTLY POSITIVE.              *
*                                                                 *
* FILES:    BELOW IS A BRIEF DESCRIPTION OF EACH FILE'S EXPECTED CONTENTS.*
*                                                                 *
*           DATA:    ON THE FIRST LINE OF THIS FILE SHOULD BE THE SAMPLE*
*                    SIZE, N.  ON THE FOLLOWING LINES SHOULD BE THE      *
*                    SAMPLE VALUES ONE PER LINE.                  *
*                                                                 *
*           PDF:     THE NORMALIZED PROBABILITY DENSITY FUNCTION  *
*                                                                 *
*           CDF:     THE CUMULATIVE DISTRIBUTION FUNCTION         *
*******************************************************************

        PROGRAM MEDLOG(DATA,PDF,CDF)

* SPECIFICATION AND DATA STATEMENTS

        INTEGER I, J, K, N
        REAL BOTTOM, F, FHAT, H, MAX, MIN, R, S, STEP, TOP, X, Z
        DIMENSION F(1000), X(1000,2), Z(1000)
        DIMENSION IRAY(6)
        DATA IRAY/6* -0/

* SET PRINT LIMIT TO ZERO, AND CALL SYSTEMC TO INHIBIT PRINTING
* OF ERROR 115

        IRAY(4)=0
        CALL SYSTEMC(115,IRAY)

* OPENING THE FILES

        OPEN(UNIT=1,FILE='DATA',STATUS='OLD')
        REWIND 1
        OPEN(UNIT=2,FILE='PDF',STATUS='NEW')
        REWIND 2
        OPEN(UNIT=3,FILE='CDF',STATUS='NEW')
        REWIND 3
```

- READING DATA FROM THE INPUT FILES

```
          WRITE(*,'("INPUT A VALUE FOR H, THE SMOOTHING PARAMETER.")')
5         READ(*,*) H
          IF (H .LE. 0) THEN
            WRITE(*,'("H MUST BE GREATER THAN 0.")')
            WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
            GOTO 5
          END IF
          READ(1,100) N
100       FORMAT(I10)
          MAX = 0.0
          MIN = 0.0
          DO 10 I=1,N
            READ(1,*) X(I,1)
            IF ( X(I,1) .GT. MAX ) THEN
              MAX = X(I,1)
            ELSE
              IF ( X(I,1) .LT. MIN) THEN
                MIN = X(I,1)
              END IF
            END IF
            X(I,2) =LOG(.5*(1.+SQRT(1.+(2.*H/X(I,1))**2)))
10        CONTINUE
```

- DETERMINING USER REQUIREMENTS

```
          WRITE(*,'("THE SAMPLE VALUES RANGE FROM")')
          WRITE(*,*) MIN
          WRITE(*,'("TO")')
          WRITE(*,*) MAX
          WRITE(*,'("INPUT THE FIRST NON-NEGATIVE POINT AT WHICH")')
          WRITE(*,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
15        READ(*,*) BOTTOM
          IF (BOTTOM .LE. 0) THEN
            WRITE(*,'("THIS VALUE MUST BE GREATER THAN ZERO.")')
            WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
            GOTO 15
          END IF
          WRITE(*,'("INPUT THE LAST POINT AT WHICH")')
          WRITE(*,'("THE DENSITY ESTIMATE WILL BE COMPUTED.")')
25        READ(*,*) TOP
          IF (TOP .LT. BOTTOM) THEN
            WRITE(*,'("THIS LAST VALUE MUST EXCEED THE FIRST VALUE.")')
            WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
            GOTO 25
          END IF
```

```
                WRITE(*,'("INPUT THE SIZE SPACING AT WHICH KERNEL DENSITY")')
                WRITE(*,'("ESTIMATES ARE DESIRED.   FOR EXAMPLE, IF THE")')
                WRITE(*,'("RANGE FOR THE SAMPLE IS 50.0 THEN A SIZE SPACING")')
                WRITE(*,'("EQUAL TO 0.10 WOULD PRODUCE DENSITY ESTIMATES")')
                WRITE(*,'("AT 0.1, 0.2, 0.3,...,50.0.)")')
  35            READ(*,*) STEP
                IF (STEP .LE. 0) THEN
                  WRITE(*,'("SIZE SPACING MUST BE GREATER THAN 0.")')
                  WRITE(*,'("PLEASE INPUT A NEW VALUE.")')
                  GOTO 35
                END IF


*  CALCULATING AND STORING F, THE NORMALIZED PROBABILITY DENSITY FUNCTION

                J = 1
                DO 20 R = BOTTOM, TOP, STEP
                  F(J) = 0.
                  DO 30 I= 1,N
                    A = EXP(-0.5*LOG(R/X(I,1))**2/X(I,2))
                    F(J) = F(J) + A/SQRT(X(I,2))
  30              CONTINUE
                  F(J) = F(J)/(SQRT(2.*3.1416)*R)/FLOAT(N)
                  Z(J) = R
                  WRITE(2,102) Z(J), F(J)
  102             FORMAT(2F15.10)
                  J = J + 1
  20            CONTINUE


*  NUMERICALLY CALCULATING FHAT, THE CUMULATIVE DISTRIBUTION FUNCTION, BY
*  USING SIMPSON'S 1/3 RULE OVER 2 SUBINTERVALS

                K = INT((TOP - BOTTOM) / STEP)
                S = ABS(Z(K-1)/K)/3.
                FHAT = 0.0
                DO 40 I= 1, K-2, 2
                  FHAT = FHAT + (S*(F(I) + (4*F(I+1)) + F(I+2)))
                  WRITE(3,102) Z(I), FHAT
  40            CONTINUE
                END
```

DISTRIBUTION LIST

PREVIOUS PAGE
IS BLANK

DISTRIBUTION LIST

No. of
Copies   Organization

1   Commander, U.S. Army Missile
     & Space Intelligence Center
     ATTN: AIAMS-YDL
     Redstone Arsenal, AL 35898-5500

1   Commander
     US Army Tank Automotive
       Command
     ATTN: AMSTA-TSL
     Warren, MI 48090

1   Director
     US Army TRADOC Systems
       Analysis Activity
     ATTN: ATAA-SL
     White Sands Missile Range,
       NM 88002

2   Commandant
     US Army Infantry School
     ATTN: ATSH-CD-CSO-OR
     Fort Benning, GA 31905

1   Commander
     US Army Development and
       Employment Agency
     ATTN: MODE-TED-SAB
     Fort Lewis, WA 98433

1   AFWL/SUL
     Kirtland AFB, NM 87117

1   Air Force Armament Laboratory
     ATTN: AFATL/DLODL
     Eglin AFB, FL 32542-5000

10   Central Intelligence Agency
     Office of Central Reference
     Dissemination Branch
     Room GE-47 HQS
     Washington, D.C.   20502

Aberdeen Proving Ground

Dir, USAMSAA
 ATTN: AMXSY-D
      AMXSY-MP, H. Cohen
Cdr, USATECOM
 ATTN: AMSTE-TO-F
Cdr, CRDC, AMCCOM
 ATTN: SMCCR-RSP-A
      SMCCR-MU
      SMCCR-SPS-IL

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. BRL Report Number_____Date of Report_____

2. Date Report Received_____

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.)_____

_____

_____

4. How specifically, is the report being used? (Information source, design data, procedure, source of ideas, etc.)_____

_____

_____

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided or efficiencies achieved, etc? If so, please elaborate._____

_____

_____

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.)

_____

_____

_____

CURRENT
ADDRESS

Name_____

Organization_____

Address_____

City, State, Zip_____

7. If indicating a Change of Address or Address Correction, please provide the New or Correct Address in Block 6 above and the Old or Incorrect address below.

OLD
ADDRESS

Name_____

Organization_____

Address_____

City, State, Zip_____

(Remove this sheet along the perforation, fold as indicated, staple or tape closed, and mail.)

- - - - - - - - - - - - - - - - FOLD HERE - - - - - - - - - - - - - - - -

Director
U.S. Army Ballistic Research Laboratory
ATTN: SLCBR-DD-T
Aberdeen Proving Ground, MD 21005-5066

|| ||||

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE. $300

BUSINESS REPLY MAIL
FIRST CLASS   PERMIT NO 12062   WASHINGTON,DC

POSTAGE WILL BE PAID BY DEPARTMENT OF THE ARMY

Director
U.S. Army Ballistic Research Laboratory
ATTN: SLCBR-DD-T
Aberdeen Proving Ground, MD 21005-9989

- - - - - - - - - - - - - - FOLD HERE - - - - - - - - - - - - - - - - -

END

FILMED

3-86

DTIC